Pathema

TIGR/BRC

Owen White

Oct, 13th 2004

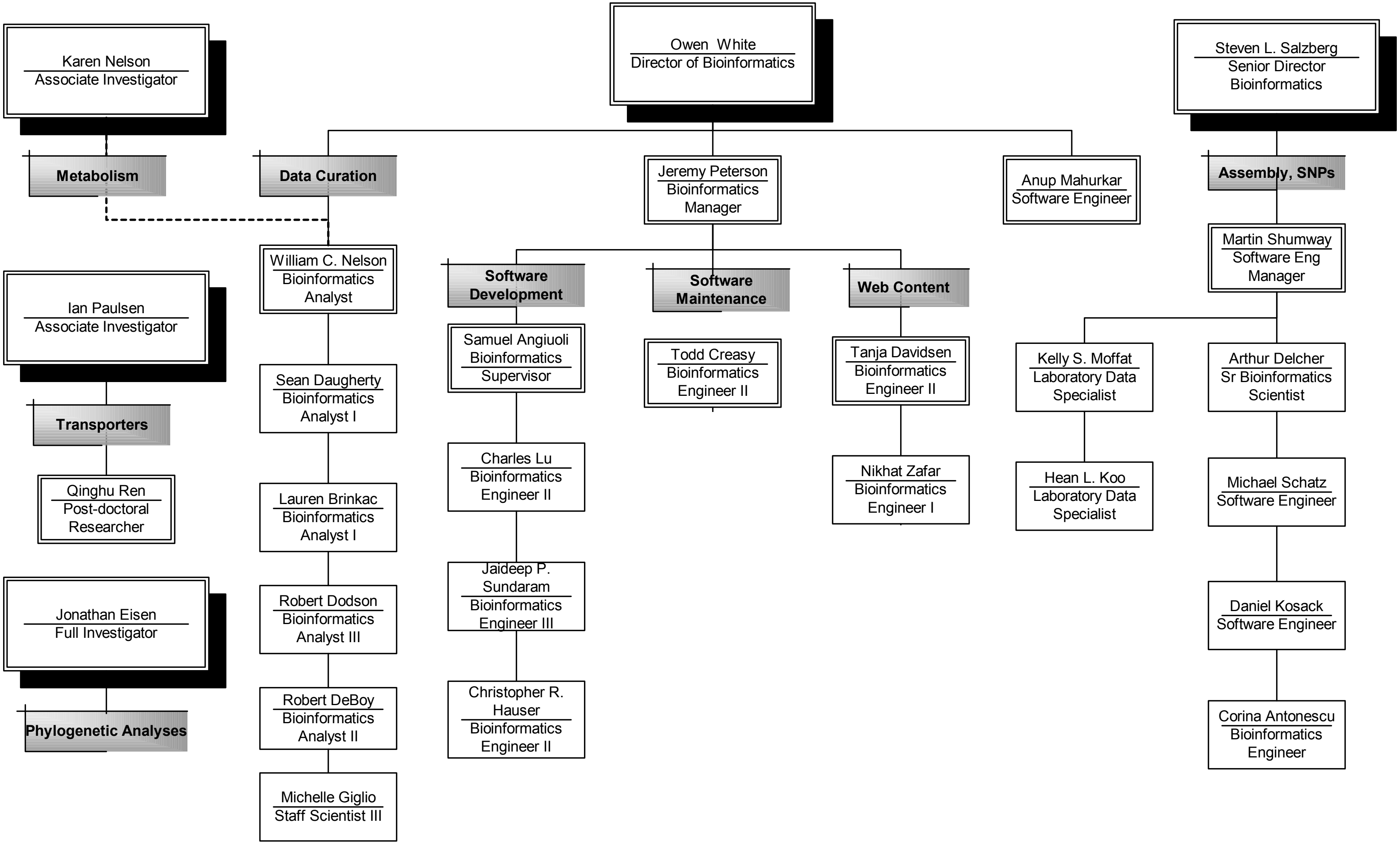
Primary Focus

- Annotation
  - Gene features/GO assignments
  - Biochemical pathways
- Comparative analysis
- SNPs across all category A-C orgs
- Instructional Classes
  - Annotation/SOPs
  - Software
  - Web usage
- Open Source Software Support

TIGR/BRC Organisms

- *Bacillus anthracis,*
- *Burkholderia mallei*
- *Burkholderia pseudomallei*
- *Clostridium botulinum*
- *Clostridium perfringens*
- *Francisella tularensis*.

People



Owen White
Director of Bioinformatics

Karen Nelson
Associate Investigator

Metabolism

Ian Paulsen
Associate Investigator

Transporters

Qinghu Ren
Post-doctoral Researcher

Jonathan Eisen
Full Investigator

Phylogenetic Analyses

Data Curation

William C. Nelson
Bioinformatics Analyst

Sean Daugherty
Bioinformatics Analyst I

Lauren Brinkac
Bioinformatics Analyst I

Robert Dodson
Bioinformatics Analyst III

Robert DeBoy
Bioinformatics Analyst II

Michelle Giglio
Staff Scientist III

+2

Jeremy Peterson
Bioinformatics Manager

Software Development

Samuel Angiuoli
Bioinformatics Supervisor

Charles Lu
Bioinformatics Engineer II

Jaideep P. Sundaram
Bioinformatics Engineer III

Christopher R. Hauser
Bioinformatics Engineer II

Software Maintenance

Todd Creasy
Bioinformatics Engineer II

Web Content

Tanja Davidsen
Bioinformatics Engineer II

Nikhat Zafar
Bioinformatics Engineer I

Anup Mahurkar
Software Engineer

Steven L. Salzberg
Senior Director Bioinformatics

Assembly, SNPs

Martin Shumway
Software Eng Manager

Kelly S. Moffat
Laboratory Data Specialist

Hean L. Koo
Laboratory Data Specialist

Arthur Delcher
Sr Bioinformatics Scientist

Michael Schatz
Software Engineer

Daniel Kosack
Software Engineer

Corina Antonescu
Bioinformatics Engineer

Microbial Software & Services

www.tigr.org/software

Software

- Mummer - whole genome alignment.
- Glimmer - gene finding system.
- Manatee - manual annotation tool.
- Workflow - custom pipelines.
- Sybil - comparative analysis system.

Mummer 3.0

- MUMs: Maximal Unique Matches
  - Algorithm finds all matches
  - String them together and align gaps
- Suffix trees
  - Fast alignment of long DNA sequences
  - Linear time and space requirements
  - Streaming algorithm
- Memory maximization
  - 2 year dev time optimizing suffix tree impl.

Mummer Performance

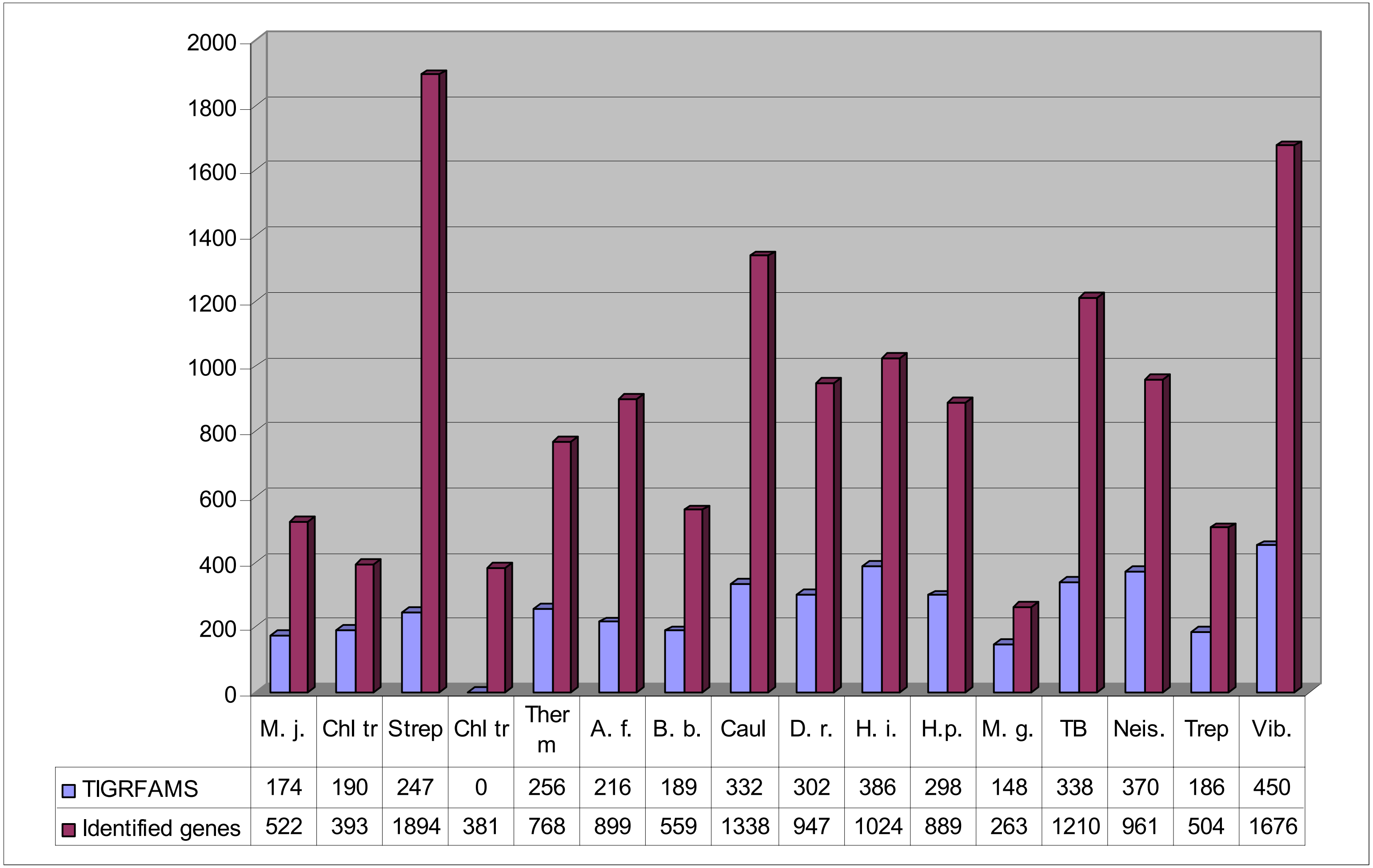
On a 2.4 GHz Pentium PC running Linux:

	3.0		2.0	
	Mb	Time	Mb	Time
E. coli K12 vs. E. coli O157	78	13s	102	14s
P. falciparum all chromosomes vs. P. yoelii	552	16:15	752	18:29
D. melanogaster arm 2L vs. D. pseudoobscura all contigs	467	13:55	465	14:39

TIGRFams

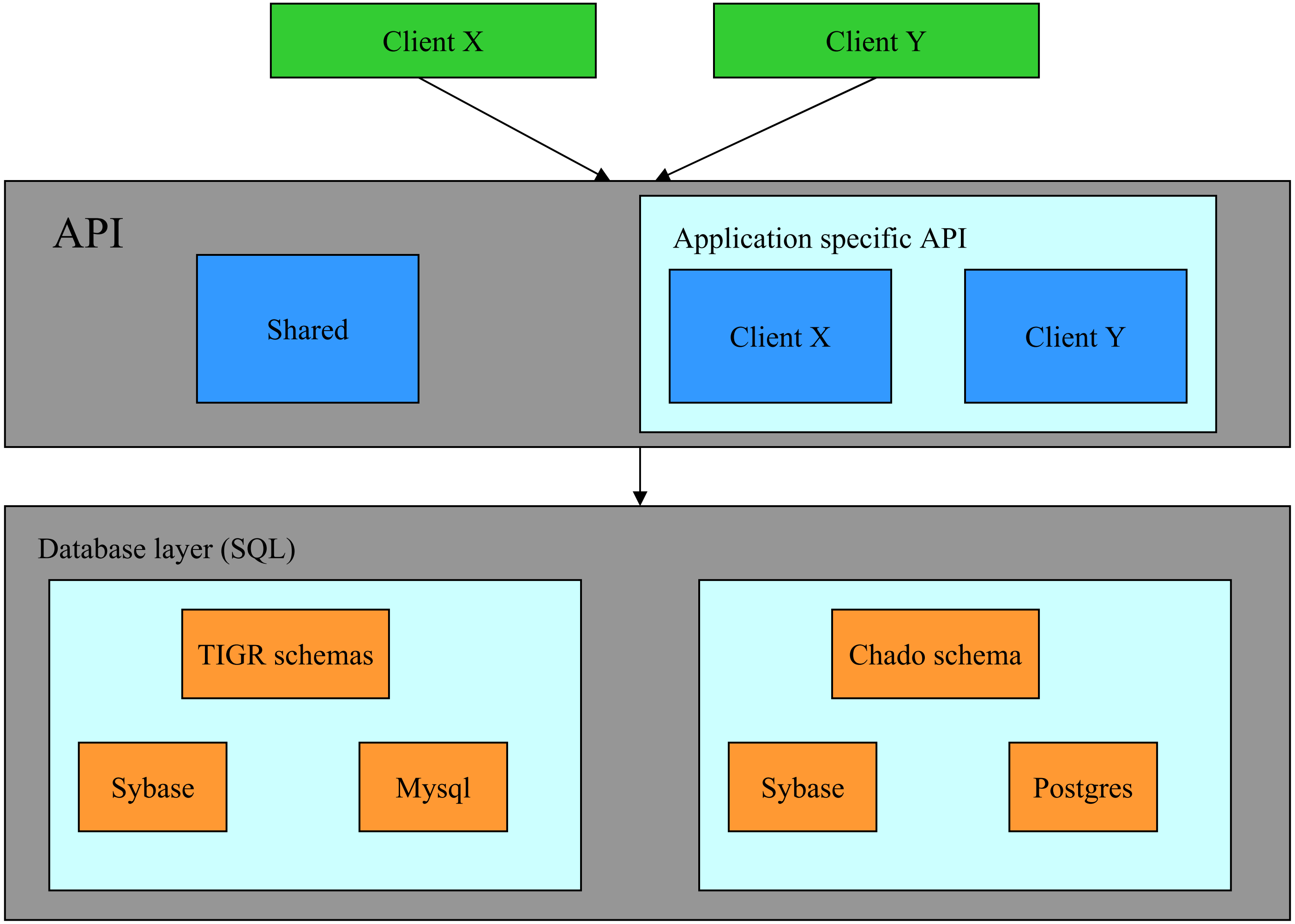
- Heavily curated multiple alignments based on protein families of the same function.
- Proposed “cure” for transitive annotation.
- Based on Hidden Markov Models (HMMs).
- >1,000 families.
- Complete assignments to GO.
- Cutoff scores for each family.
  - Trusted (automated name assignment)
  - Noise (manual inspection required)
- Downloadable. Fully integrated into Interpro.

TIGRFAMs: **Orthologous Families**



	M. j.	Chl tr	Strep	Chl tr	Therm	A. f.	B. b.	Caul	D. r.	H. i.	H.p.	M. g.	TB	Neis.	Trep	Vib.
🟦 TIGRFAMS	174	190	247	0	256	216	189	332	302	386	298	148	338	370	186	450
🟥 Identified genes	522	393	1894	381	768	899	559	1338	947	1024	889	263	1210	961	504	1676

3-tier Architecture

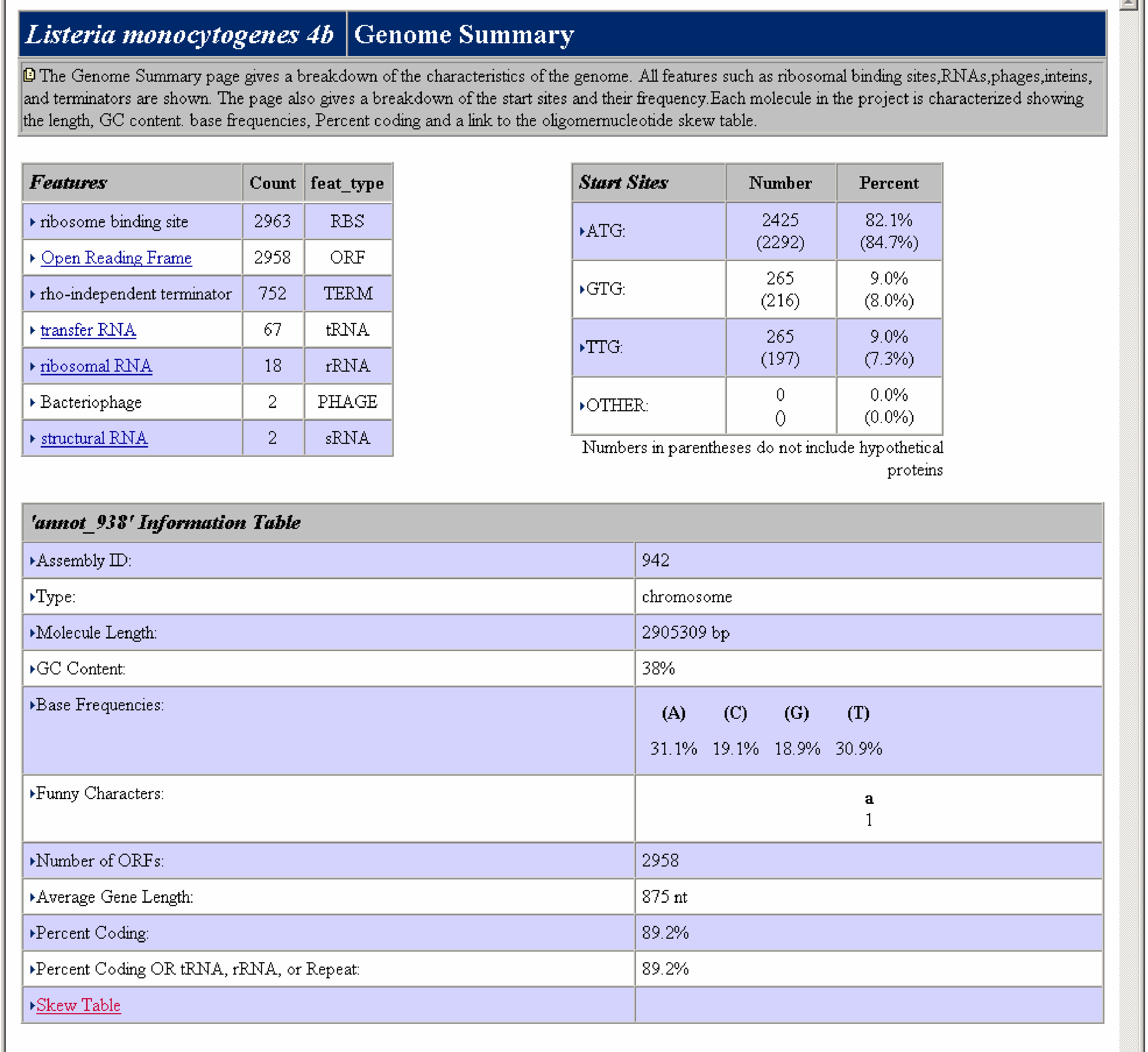


3-tier Architecture

- API
  - Unified access to data for client apps
  - Decouples client applications from SQL
- Portable
  - Cross schema/vendor support
  - Perl
- Extensible
  - Application specific API and database layer
- Thin/Lightweight
  - Single layer between client and SQL

Manatee

- MANual Annotation Tool Etc, Etc…



Listeria monocytogenes 4b Genome Summary

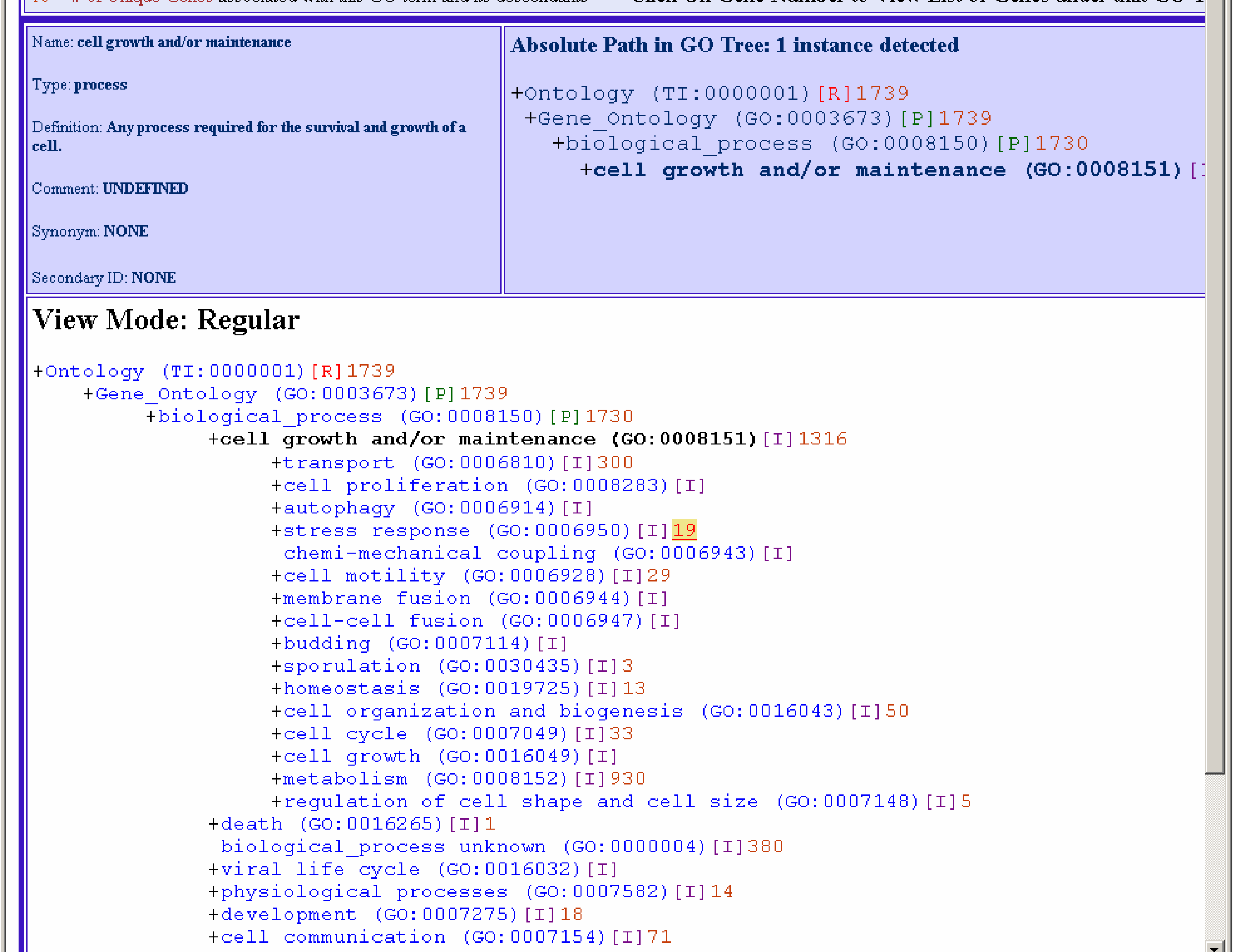
The Genome Summary page gives a breakdown of the characteristics of the genome. All features such as ribosomal binding sites,RNAs,phages,inteins, and terminators are shown. The page also gives a breakdown of the start sites and their frequency.Each molecule in the project is characterized showing the length, GC content. base frequencies, Percent coding and a link to the oligomernucleotide skew table.

Features	**Count**	**feat_type**
ribosome binding site	2963	RBS
Open Reading Frame	2958	ORF
rho-independent terminator	752	TERM
transfer RNA	67	tRNA
ribosomal RNA	18	rRNA
Bacteriophage	2	PHAGE
structural RNA	2	sRNA

Start Sites	**Number**	**Percent**
ATG:	2425 (2292)	82.1% (84.7%)
GTG:	265 (216)	9.0% (8.0%)
TTG:	265 (197)	9.0% (7.3%)
OTHER:	0 0	0.0% (0.0%)

Numbers in parentheses do not include hypothetical proteins

'annot_938' Information Table	
Assembly ID:	942
Type:	chromosome
Molecule Length:	2905309 bp
GC Content:	38%
Base Frequencies:	**(A)** **(C)** **(G)** **(T)** 31.1% 19.1% 18.9% 30.9%
Funny Characters:	**a** 1
Number of ORFs:	2958
Average Gene Length:	875 nt
Percent Coding:	89.2%
Percent Coding OR tRNA, rRNA, or Repeat:	89.2%
Skew Table	



Name: **cell growth and/or maintenance**

Type: **process**

Definition: **Any process required for the survival and growth of a cell.**

Comment: **UNDEFINED**

Synonym: **NONE**

Secondary ID: **NONE**

Absolute Path in GO Tree: 1 instance detected

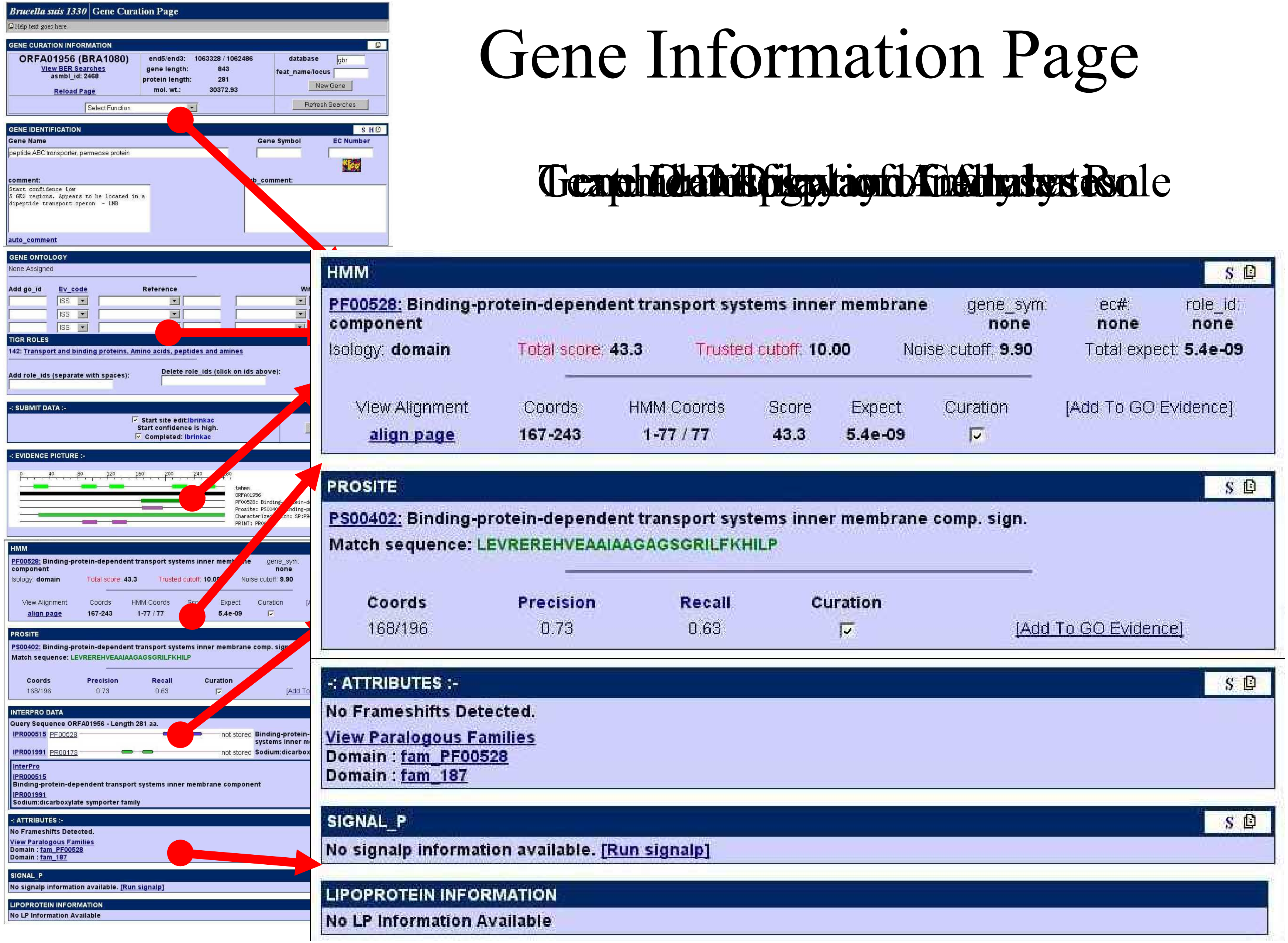
```
+Ontology (TI:0000001)[R]1739
 +Gene_Ontology (GO:0003673)[P]1739
   +biological_process (GO:0008150)[P]1730
     +cell growth and/or maintenance (GO:0008151)[
```

View Mode: Regular

```
+Ontology (TI:0000001)[R]1739
    +Gene_Ontology (GO:0003673)[P]1739
         +biological_process (GO:0008150)[P]1730
              +cell growth and/or maintenance (GO:0008151)[I]1316
                   +transport (GO:0006810)[I]300
                   +cell proliferation (GO:0008283)[I]
                   +autophagy (GO:0006914)[I]
                   +stress response (GO:0006950)[I]19
                    chemi-mechanical coupling (GO:0006943)[I]
                   +cell motility (GO:0006928)[I]29
                   +membrane fusion (GO:0006944)[I]
                   +cell-cell fusion (GO:0006947)[I]
                   +budding (GO:0007114)[I]
                   +sporulation (GO:0030435)[I]3
                   +homeostasis (GO:0019725)[I]13
                   +cell organization and biogenesis (GO:0016043)[I]50
                   +cell cycle (GO:0007049)[I]33
                   +cell growth (GO:0016049)[I]
                   +metabolism (GO:0008152)[I]930
                   +regulation of cell shape and cell size (GO:0007148)[I]5
              +death (GO:0016265)[I]1
               biological_process unknown (GO:0000004)[I]380
              +viral life cycle (GO:0016032)[I]
              +physiological processes (GO:0007582)[I]14
              +development (GO:0007275)[I]18
              +cell communication (GO:0007154)[I]71
```

Gene Information Page

[illegible]



Brucella suis 1330 **Gene Curation Page**

Help text goes here.

GENE CURATION INFORMATION

ORFA01956 (BRA1080) View BER Searches asmbl_id: 2468 Reload Page	end5/end3: gene length: protein length: mol. wt.:	1063328 / 1062486 843 281 30372.93	database feat_name/locus	gbr

Select Function | New Gene | Refresh Searches

GENE IDENTIFICATION

Gene Name	Gene Symbol	EC Number
peptide ABC transporter, permease protein		

comment:

```
Start confidence Low
5 GES regions. Appears to be located in a
dipeptide transport operon  - LMB
```

ub_comment:

auto_comment

GENE ONTOLOGY

None Assigned

Add go_id	Ev_code	Reference		Wi
	ISS			
	ISS			
	ISS			

TIGR ROLES

142: Transport and binding proteins, Amino acids, peptides and amines

Add role_ids (separate with spaces):

Delete role_ids (click on ids above):

-: SUBMIT DATA :-

☑ Start site edit:lbrinkac
Start confidence is high.
☑ Completed: lbrinkac

-: EVIDENCE PICTURE :-

0
40
80
120
160
200
240
280
tmhmm
ORFA01956
PF00528: Binding-protein-de
Prosite: PS0040 Binding-pr
Characterized Match: SP:P9
PRINT: PR00

HMM

PF00528: Binding-protein-dependent transport systems inner membrane component | gene_sym: none

Isology: domain | Total score: 43.3 | Trusted cutoff: 10.00 | Noise cutoff: 9.90

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[A
align page	167-243	1-77 / 77		5.4e-09	☑	

PROSITE

PS00402: Binding-protein-dependent transport systems inner membrane comp. sign

Match sequence: LEVREREHVEAAIAAGAGSGRILFKHILP

Coords	Precision	Recall	Curation	
168/196	0.73	0.63	☑	[Add To

INTERPRO DATA

Query Sequence ORFA01956 - Length 281 aa.

IPR000515	PF00528	not stored	Binding-protein- systems inner m
IPR001991	PR00173	not stored	Sodium:dicarbox

InterPro

IPR000515
Binding-protein-dependent transport systems inner membrane component

IPR001991
Sodium:dicarboxylate symporter family

-: ATTRIBUTES :-

No Frameshifts Detected.

View Paralogous Families
Domain : fam_PF00528
Domain : fam_187

SIGNAL_P

No signalp information available. [Run signalp]

LIPOPROTEIN INFORMATION

No LP Information Available

HMM

PF00528: Binding-protein-dependent transport systems inner membrane component | gene_sym: none | ec#: none | role_id: none

Isology: domain | Total score: 43.3 | Trusted cutoff: 10.00 | Noise cutoff: 9.90 | Total expect: 5.4e-09

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
align page	167-243	1-77 / 77	43.3	5.4e-09	☑	

PROSITE

PS00402: Binding-protein-dependent transport systems inner membrane comp. sign.

Match sequence: LEVREREHVEAAIAAGAGSGRILFKHILP

Coords	Precision	Recall	Curation	
168/196	0.73	0.63	☑	[Add To GO Evidence]

-: ATTRIBUTES :-

No Frameshifts Detected.

View Paralogous Families
Domain : fam_PF00528
Domain : fam_187

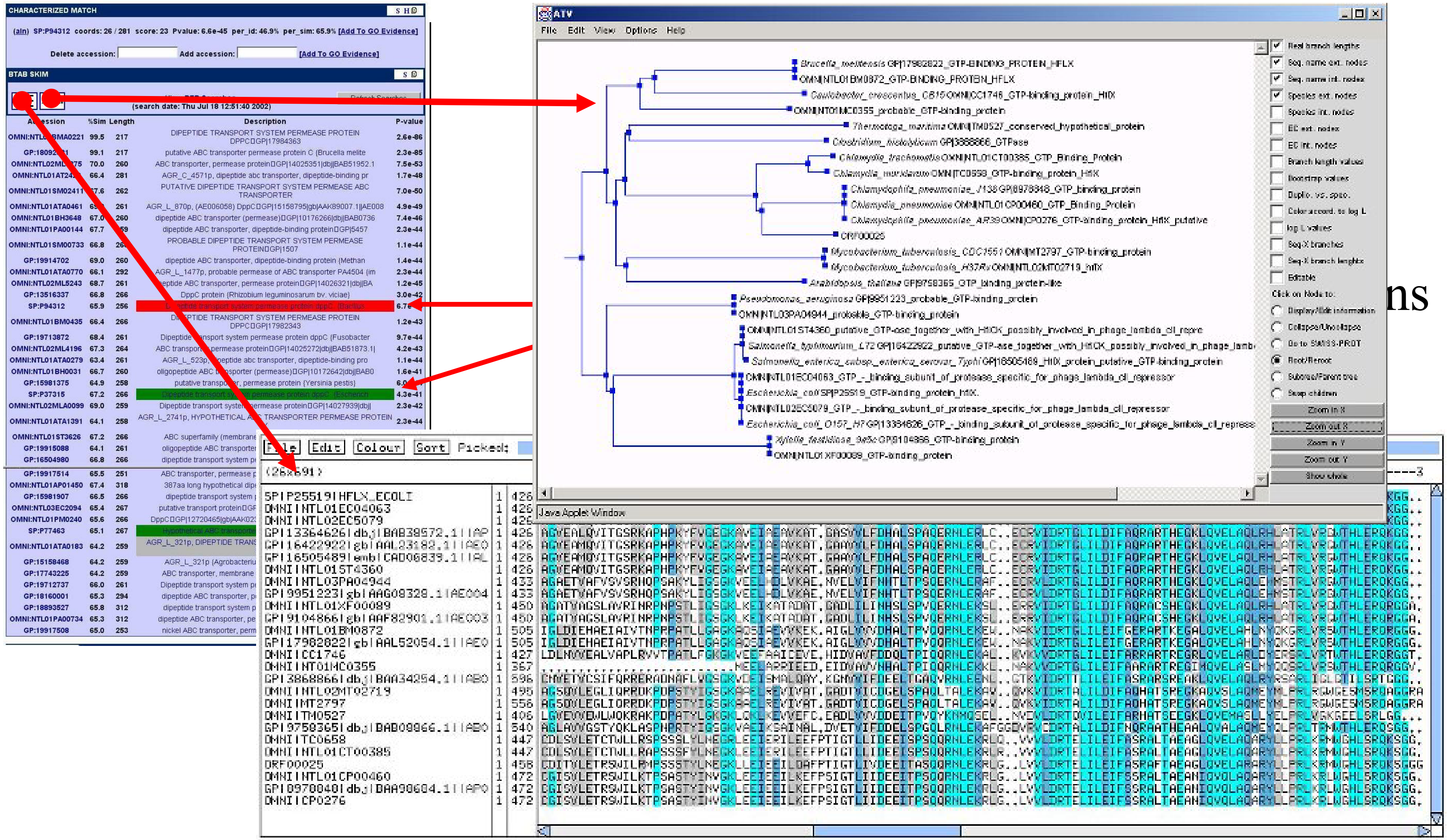
SIGNAL_P

No signalp information available. [Run signalp]

LIPOPROTEIN INFORMATION

No LP Information Available

Pair-wise Alignment Summary



CHARACTERIZED MATCH

(aln) SP:P94312 coords: 26 / 281 score: 23 Pvalue: 6.6e-45 per_id: 46.9% per_sim: 65.9% [Add To GO Evidence]

Delete accession: Add accession: [Add To GO Evidence]

BTAB SKIM

(search date: Thu Jul 18 12:51:40 2002)

Accession	%Sim	Length	Description	P-value
OMNI:NTL01BMA0221	99.5	217	DIPEPTIDE TRANSPORT SYSTEM PERMEASE PROTEIN DPPC□GP|17984363	2.6e-86
GP:18092[illegible]1	99.1	217	putative ABC transporter permease protein C {Brucella melite	2.3e-85
OMNI:NTL02ML[illegible]75	70.0	260	ABC transporter, permease protein□GP|14025351|dbj|BAB51952.1	7.5e-53
OMNI:NTL01AT24[illegible]	66.4	281	AGR_C_4571p, dipeptide abc transporter, dipeptide-binding pr	1.7e-48
OMNI:NTL01SM02411	67.6	262	PUTATIVE DIPEPTIDE TRANSPORT SYSTEM PERMEASE ABC TRANSPORTER	7.0e-50
OMNI:NTL01ATA0461	69.[illegible]	261	AGR_L_870p, (AE006058) DppC□GP|15158795|gb|AAK89007.1||AE008	4.9e-49
OMNI:NTL01BH3648	67.0	260	dipeptide ABC transporter (permease)□GP|10176266|dbj|BAB0736	7.4e-46
OMNI:NTL01PA00144	67.7	[illegible]59	dipeptide ABC transporter, dipeptide-binding protein□GP|5457	2.3e-44
OMNI:NTL01SM00733	66.8	26[illegible]	PROBABLE DIPEPTIDE TRANSPORT SYSTEM PERMEASE PROTEIN□GP|1507	1.1e-44
GP:19914702	69.0	260	dipeptide ABC transporter, dipeptide-binding protein {Methan	1.4e-44
OMNI:NTL01ATA0770	66.1	292	AGR_L_1477p, probable permease of ABC transporter PA4504 {im	2.3e-44
OMNI:NTL02ML5243	68.7	261	[illegible]eptide ABC transporter, permease protein□GP|14026321|dbj|BA	1.2e-45
GP:13516337	66.8	266	DppC protein {Rhizobium leguminosarum bv. viciae}	3.0e-42
SP:P94312	65.9	256	Dipeptide transport system permease protein dppC {Bacillus	6.7e[illegible]
OMNI:NTL01BM0435	66.4	266	DIPEPTIDE TRANSPORT SYSTEM PERMEASE PROTEIN DPPC□GP|17982343	1.2e-43
GP:19713872	68.4	261	Dipeptide transport system permease protein dppC {Fusobacter	9.7e-44
OMNI:NTL02ML4196	67.3	264	ABC transporter, permease protein□GP|14025272|dbj|BAB51873.1|	4.2e-43
OMNI:NTL01ATA0279	63.4	261	AGR_L_523p, dipeptide abc transporter, dipeptide-binding pro	1.1e-44
OMNI:NTL01BH0031	66.7	260	oligopeptide ABC transporter (permease)□GP|10172642|dbj|BAB0	1.6e-41
GP:15981375	64.9	258	putative transporter, permease protein {Yersinia pestis}	6.0[illegible]
SP:P37315	67.2	266	Dipeptide transport system permease protein dppC {Escherich	4.3e-41
OMNI:NTL02MLA0099	69.0	259	Dipeptide transport system permease protein□GP|14027939|dbj|	2.3e-42
OMNI:NTL01ATA1391	64.1	258	AGR_L_2741p, HYPOTHETICAL ABC TRANSPORTER PERMEASE PROTEIN	2.3e-44
OMNI:NTL01ST3626	67.2	266	ABC superfamily (membrane	
GP:19915088	64.1	261	oligopeptide ABC transporter	
GP:16504980	66.8	266	dipeptide transport system pe	
GP:19917514	65.5	251	ABC transporter, permease p	
OMNI:NTL01AP01450	67.4	318	387aa long hypothetical dipe	
GP:15981907	66.5	266	dipeptide transport system p	
OMNI:NTL03EC2094	65.4	267	putative transport protein□GP	
OMNI:NTL01PM0240	65.6	266	DppC□GP|12720465|gb|AAK023	
SP:P77463	65.1	267	Hypothetical ABC transporte	
OMNI:NTL01ATA0183	64.2	259	AGR_L_321p, DIPEPTIDE TRANS	
GP:15158468	64.2	259	AGR_L_321p {Agrobacteriu	
GP:17743225	64.2	259	ABC transporter, membrane	
GP:19712737	66.0	261	Dipeptide transport system pe	
GP:18160001	65.3	294	dipeptide ABC transporter, p	
GP:18893527	65.8	312	dipeptide transport system p	
OMNI:NTL01PA00734	65.3	312	dipeptide ABC transporter, pe	
GP:19917508	65.0	253	nickel ABC transporter, perm	

ATV

Brucella_melitensis GP|17982822_GTP-BINDING_PROTEIN_HFLX
OMNI|NTL01BM0872_GTP-BINDING_PROTEIN_HFLX
Caulobacter_crescentus_CB15 OMNI|CC1746_GTP-binding_protein_HflX
OMNI|NT01MC0355_probable_GTP-binding_protein
Thermotoga_maritima OMNI|TM0527_conserved_hypothetical_protein
Clostridium_histolyticum GP|3868866_GTPase
Chlamydia_trachomatis OMNI|NTL01CT00385_GTP_Binding_Protein
Chlamydia_muridarum OMNI|TC0658_GTP-binding_protein_HflX
Chlamydophila_pneumoniae_J138 GP|8978848_GTP_binding_protein
Chlamydia_pneumoniae OMNI|NTL01CP00460_GTP_Binding_Protein
Chlamydophila_pneumoniae_AR39 OMNI|CP0276_GTP-binding_protein_HflX_putative
ORF00025
Mycobacterium_tuberculosis_CDC1551 OMNI|MT2797_GTP-binding_protein
Mycobacterium_tuberculosis_H37Rv OMNI|NTL02MT02719_hflX
Arabidopsis_thaliana GP|9758365_GTP_binding_protein-like
Pseudomonas_aeruginosa GP|9951223_probable_GTP-binding_protein
OMNI|NTL03PA04944_probable_GTP-binding_protein
OMNI|NTL01ST4360_putative_GTP-ase_together_with_HflCK_possibly_involved_in_phage_lambda_cII_repre
Salmonella_typhimurium_LT2 GP|16422922_putative_GTP-ase_together_with_HflCK_possibly_involved_in_phage_lamb
Salmonella_enterica_subsp._enterica_serovar_Typhi GP|16505489_HflX_protein_putative_GTP-binding_protein
OMNI|NTL01EC04063_GTP_-_binding_subunit_of_protease_specific_for_phage_lambda_cII_repressor
Escherichia_coli SP|P25519_GTP-binding_protein_hflX.
OMNI|NTL02EC5079_GTP_-_binding_subunit_of_protease_specific_for_phage_lambda_cII_repressor
Escherichia_coli_O157_H7 GP|13364626_GTP_-_binding_subunit_of_protease_specific_for_phage_lambda_cII_repress
Xylella_fastidiosa_9a5c GP|9104866_GTP-binding_protein
OMNI|NTL01XF00089_GTP-binding_protein

☑ Real branch lengths
☑ Seq. name ext. nodes
☑ Seq. name int. nodes
☑ Species ext. nodes
☐ Species int. nodes
☐ EC ext. nodes
☐ EC int. nodes
☐ Branch length values
☐ Bootstrap values
☐ Duplic. vs. spec.
☐ Color accord. to log L
☐ log L values
☐ Seq-X branches
☐ Seq-X branch lenghts
☐ Editable

Click on Node to:
○ Display/Edit information
○ Collapse/Uncollapse
○ Go to SWISS-PROT
● Root/Reroot
○ Subtree/Parent tree
○ Swap children

Java Applet Window

(26x591)

----3

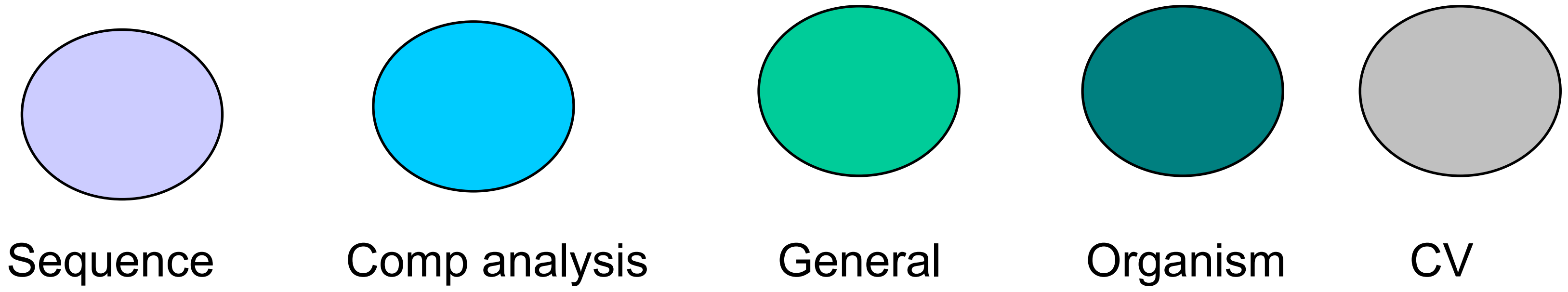
```
SP|P25519|HFLX_ECOLI            1  426  ...KGG..
OMNI|NTL01EC04063               1  426  ...KGG..
OMNI|NTL02EC5079                1  426  ...KGG..
GP|13364626|dbj|BAB38572.1||AP  1  426  AGVEALQVITGSRKAPHPKYFVGEGKAVEIAEAVKAT.GASVVLFDHALSPAQERNLERLC..ECRVIDRTGLILDIFAQRARTHEGKLQVELAQLRHLATRLVRGWTHLERQKGG..
GP|16422922|gb|AAL23182.1||AE0  1  426  AGVEAMQVITGSRKAPHPKYFVGEGKAVEIAEAVKAT.GAAVVLFDHALSPAQERNLERLC..ECRVIDRTGLILDIFAQRARTHEGKLQVELAQLRHLATRLVRGWTHLERQKGG..
GP|16505489|emb|CAD06839.1||AL  1  426  AGVEAMQVITGSRKAPHPKYFVGEGKAVEIAEAVKAT.GAAVVLFDHALSPAQERNLERLC..ECRVIDRTGLILDIFAQRARTHEGKLQVELAQLRHLATRLVRGWTHLERQKGG..
OMNI|NTL01ST4360                1  426  AGVEAMQVITGSRKAPHPKYFVGEGKAVEIAEAVKAT.GAAVVLFDHALSPAQERNLERLC..ECRVIDRTGLILDIFAQRARTHEGKLQVELAQLRHLATRLVRGWTHLERQKGG..
OMNI|NTL03PA04944               1  433  AGAETVAFVSVSRHQPSAKYLIGSGKVEELHDLVKAE.NVELVIFNHTLTPSQERNLERAF..ECRVLDRTGLILDIFAQRARTHEGKLQVELAQLEHMSTRLVRGWTHLERQKGG..
GP|9951223|gb|AAG08328.1|AE004  1  433  AGAETVAFVSVSRHQPSAKYLIGSGKVEELHDLVKAE.NVELVIFNHTLTPSQERNLERAF..ECRVLDRTGLILDIFAQRARTHEGKLQVELAQLEHMSTRLVRGWTHLERQKGG..
OMNI|NTL01XF00089               1  450  AGATVAGSLAVRINRPNPSTLIGSGKLKEIKATADAT.GADLILINHSLSPVQERNLEKSL..ERRVIDRTGLILDIFAQRACSHEGKLQVELAQLRHLATRLVRGWTHLERQRGGA.
GP|9104866|gb|AAF82901.1|AE003  1  450  AGATVAGSLAVRINRPNPSTLIGSGKLKEIKATADAT.GADLILINHSLSPVQERNLEKSL..ERRVIDRTGLILDIFAQRACSHEGKLQVELAQLRHLATRLVRGWTHLERQRGGA.
OMNI|NTL01BM0872                1  505  IGLDIEHAEIAIVTNPRPATLLGAGKADSIAEVVKEK.AIGLVVVDHALTPVQQRNLEKEW..NAKVIDRTGLILEIFGERARTKEGALQVELAHLNYQKGRLVRSWTHLERQRGGG.
GP|17982822|gb|AAL52054.1||AE0  1  505  IGLDIEHAEIAIVTNPRPATLLGAGKADSIAEVVKEK.AIGLVVVDHALTPVQQRNLEKEW..NAKVIDRTGLILEIFGERARTKEGALQVELAHLNYQKGRLVRSWTHLERQRGGG.
OMNI|CC1746                     1  427  LDLNVVEALVAPLRVVTPATLFGKGKVEEFAAICEVE.HIDVAVFDDQLTPIQQRNLEKAL..KVKVVDRTGLILEIFARRARTREGRLQVELARLDYERSRLVRTWTHLERQRGGT.
OMNI|NT01MC0355                 1  367  ..........................NEELARRIEED.EIDVAVVNHALTPIQQRNLEKKL..NAKVVDRTGLILEIFAARARTREGIMQVELASLNYQQSRLVRSWTHLERQRGGV.
GP|3868866|dbj|BAA34254.1||AB0  1  596  CNVETYCSIFQRRERADNAFLVGSGKVDEIEMALQAY.KGNVVIFDEELTGAQVRNLEENL..GTKVIDRTTLILEIFASRARSREAKLQVELAQLRYRSARLIGLGTILSRTGGG..
OMNI|NTL02MT02719               1  495  AGSDVLEGLIQRRDKPDPSTYIGSGKAAELREVIVAT.GADTVICDGELSPAQLTALEKAV..QVKVIDRTALILDIFAQHATSREGKAQVSLAQMEYMLPRLRGWGESMSRQAGGRA
OMNI|MT2797                     1  556  AGSDVLEGLIQRRDKPDPSTYIGSGKAAELREVIVAT.GADTVICDGELSPAQLTALEKAV..QVKVIDRTALILDIFAQHATSREGKAQVSLAQMEYMLPRLRGWGESMSRDAGGRA
OMNI|TM0527                     1  406  LGVEVVDWLWQKRAKPDPATYLGKGKLQKLKEVVEFC.EADLVVVDDEITPVQYKNMQSEL..NVIVLDRTQVILEIFARHATSEEGKLQVEMASLLYELPRLVGKGEELSRLGG...
GP|9758365|dbj|BAB08866.1||AB0  1  540  AGLAVVGSTYQKLASPNPRTYIGSGKVAEIKSAINAL.DVETVIFDDELSPGQLRNLEKAFGGDVRVCDRTALILDIFNQRAATHEAALQVALAQMEYQLPRLTRMWTHLERQSGG..
OMNI|TC0658                     1  447  CDLSVLETCTWLLRSPSSSLYLNEGALEEIERILEEFPTIGTLLIDEEISPSQQRNLEKRLQ..VVVLDRTELILEIFASRALTAEAGLQVELAQARYLLPRLKRMWGHLSRQKSGG.
OMNI|NTL01CT00385               1  447  CDLSVLETCTWLLRAPSSSFYLNEGKLEEIERILEEFPTIGTLLIDEEISPSQQRNLEKRLR..VVVLDRTELILEIFASRALTAEAGLQVELAQARYLLPRLKRMWGHLSRQKSGG.
ORF00025                        1  458  CDITVLETRSWILRMPSSSTYLNEGKLLEIEEILDAFPTIGTLIVDEEITASQQRNLEKRLG..LVVLDRTELILEIFASRAFTAEAGLQVELARARYLLPRLKRMWGHLSRQKSGGG
OMNI|NTL01CP00460               1  472  CGISVLETRSWILKTPSASTYINVGKLEEIEEILKEFPSIGTLIIDEEITPSQQRNLEKRLG..LVVLDRTELILEIFSSRALTAEANIQVQLAQARYLLPRLKRLWGHLSRQKSGG.
GP|8970040|dbj|BAA98604.1||AP0  1  472  CGISVLETRSWILKTPSASTYINVGKLEEIEEILKEFPSIGTLIIDEEITPSQQRNLEKRLG..LVVLDRTELILEIFSSRALTAEANIQVQLAQARYLLPRLKRLWGHLSRQKSGG.
OMNI|CP0276                     1  472  CGISVLETRSWILKTPSASTYINVGKLEEIEEILKEFPSIGTLIIDEEITPSQQRNLEKRLG..LVVLDRTELILEIFSSRALTAEANIQVQLAQARYLLPRLKRLWGHLSRQKSGG.
```

ns

Manatee: Implementation

- ~15 reports
- Linux OS
- Fully database independent.
  - Mysql
  - Sybase
- Perl Application programmer interface (API).
- Web compatible.
- Documented
  - Example databases available.
  - Online user docs.
  - Programmer/interface.
  - Installation.
- Installed at several sites/contributions from other developers received

Chado: Open Source Database



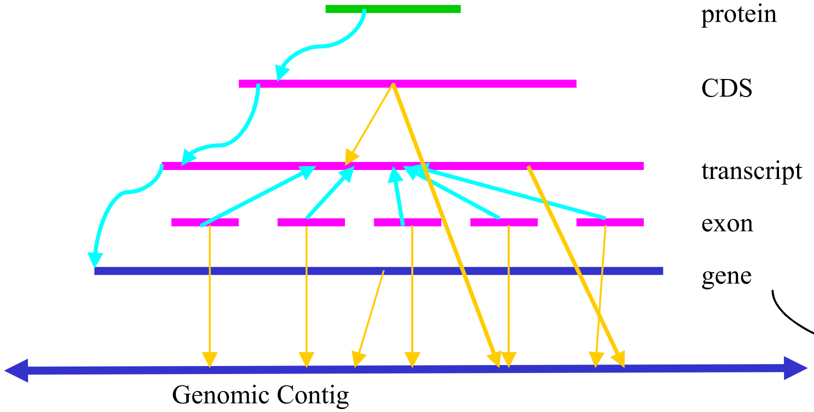
Sequence Comp analysis General Organism CV

- Collaborative relational database ~15 people.
  - Flybase/Harvard
  - GMOD Consortium
- Composed of several modules
- Freely available as open source
- Many support tools under development by several laboratories. (see www.gmod.org)

Annotation attributes

Annot. Attribute terms

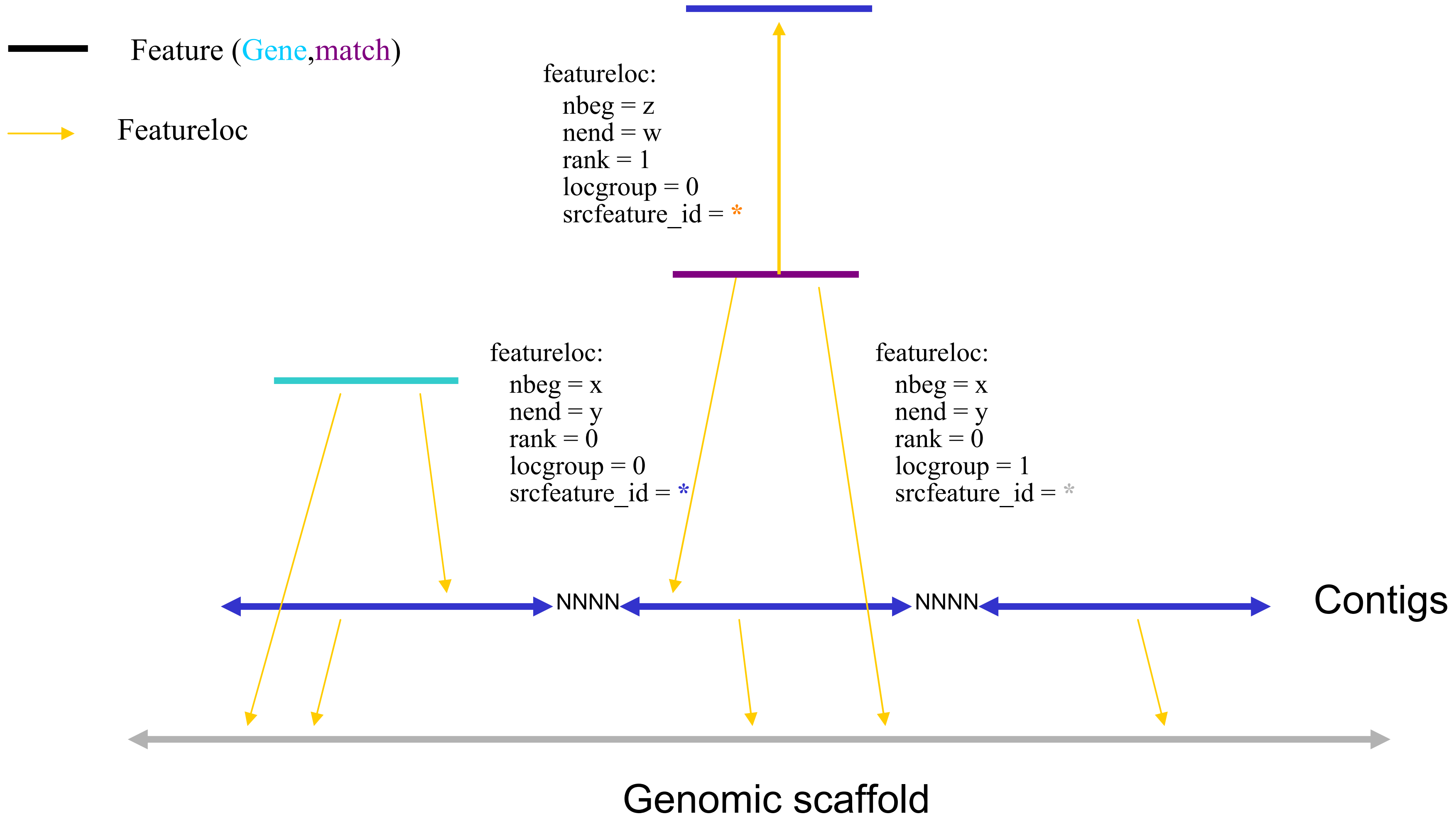
Gene name
EC#
Gene sym
GO
…



featureprop:
type_id =
value =

cvterm:
cv_id
cvterm_id
name

Scaffold mappings



Feature (Gene,match)

Featureloc

featureloc:
nbeg = z
nend = w
rank = 1
locgroup = 0
srcfeature_id = *

featureloc:
nbeg = x
nend = y
rank = 0
locgroup = 0
srcfeature_id = *

featureloc:
nbeg = x
nend = y
rank = 0
locgroup = 1
srcfeature_id = *

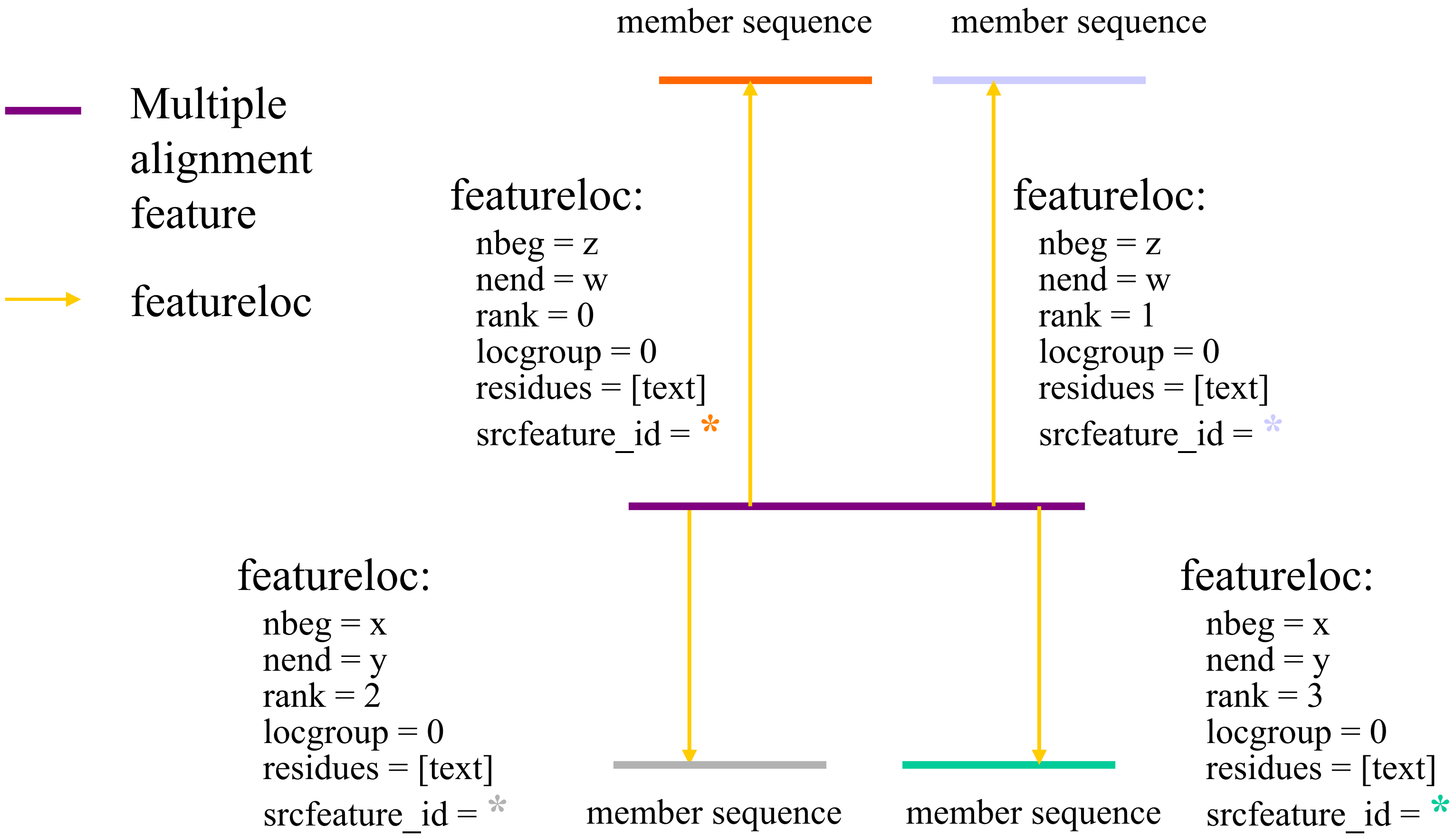
NNNN

NNNN

Contigs

Genomic scaffold

Multiple alignments



Multiple alignment feature

featureloc

member sequence

member sequence

featureloc:
nbeg = z
nend = w
rank = 0
locgroup = 0
residues = [text]
srcfeature_id = *

featureloc:
nbeg = z
nend = w
rank = 1
locgroup = 0
residues = [text]
srcfeature_id = *

featureloc:
nbeg = x
nend = y
rank = 2
locgroup = 0
residues = [text]
srcfeature_id = *

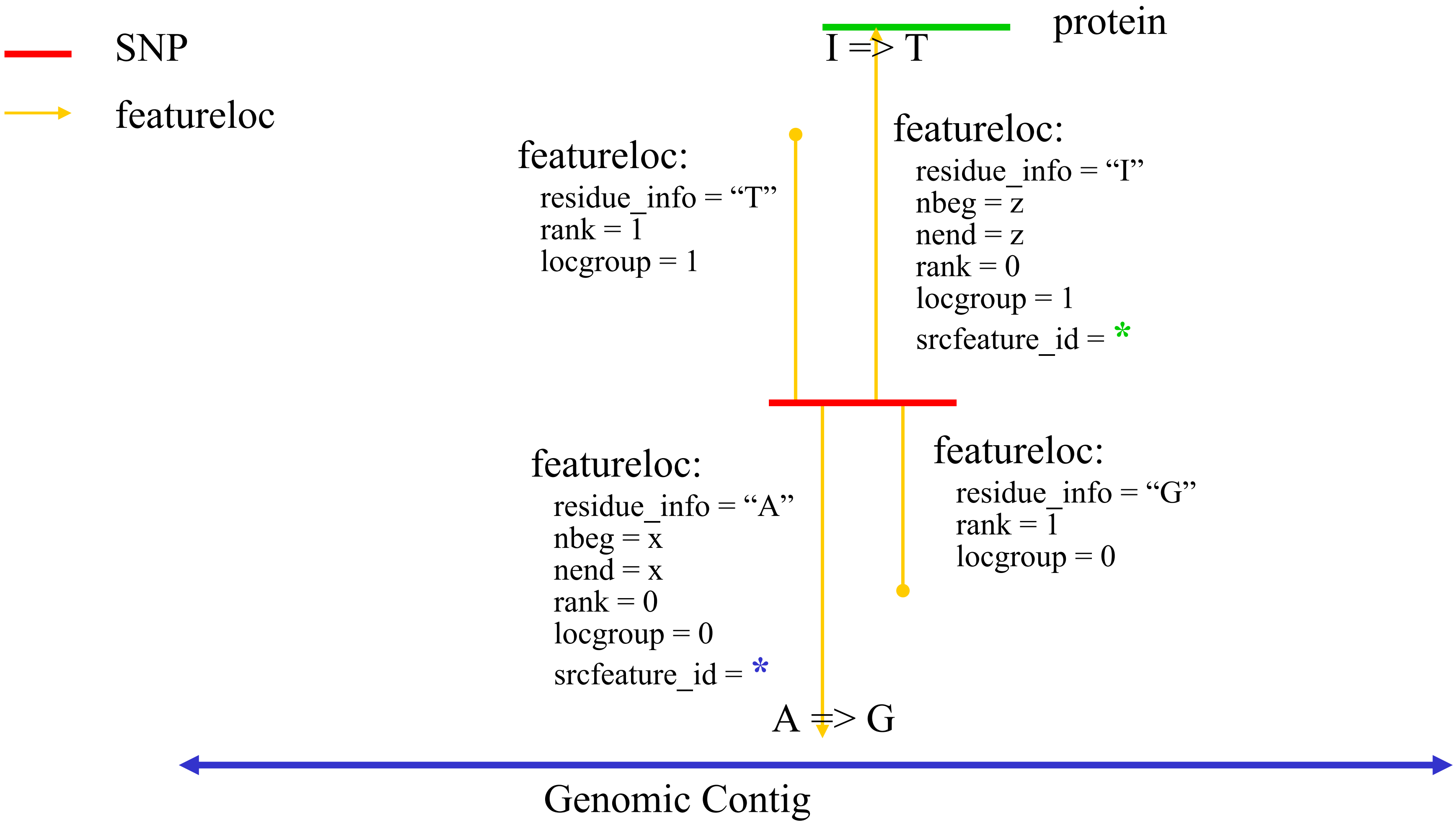
member sequence

member sequence

featureloc:
nbeg = x
nend = y
rank = 3
locgroup = 0
residues = [text]
srcfeature_id = *

Sequence variations

SNPs (redundant mapping to protein)



— SNP

→ featureloc

protein

I => T

featureloc:
residue_info = “T”
rank = 1
locgroup = 1

featureloc:
residue_info = “I”
nbeg = z
nend = z
rank = 0
locgroup = 1
srcfeature_id = *

featureloc:
residue_info = “A”
nbeg = x
nend = x
rank = 0
locgroup = 0
srcfeature_id = *

featureloc:
residue_info = “G”
rank = 1
locgroup = 0

A => G

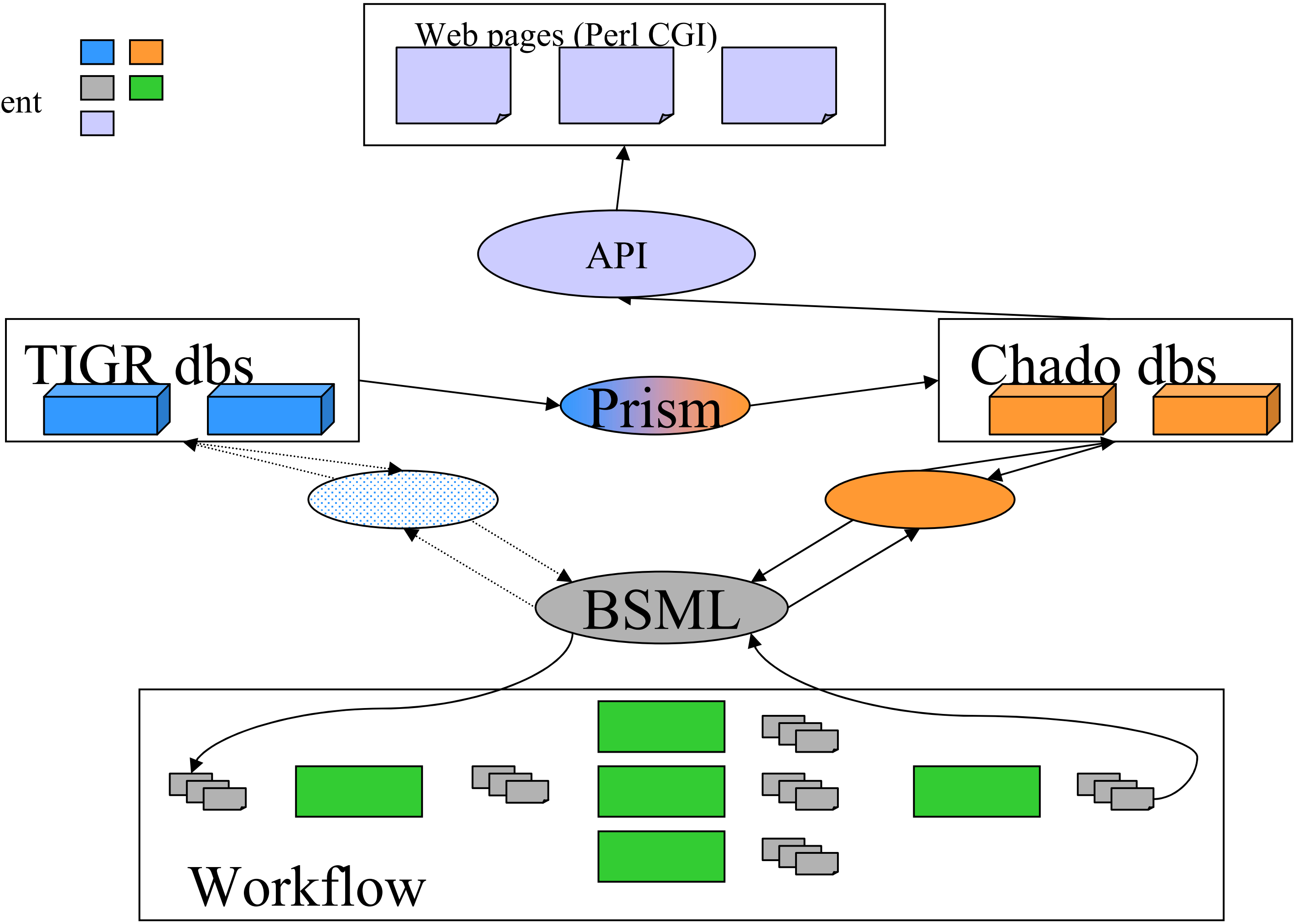
Genomic Contig

Chado ←→BSML

- No problem.
- Heavily documented
- Scaffolds
- Attribution
- CV terms
- Linkage into other documents
- Large-scale file management

Project architecture

*Database 🟦 🟧
*Data management ⬜ 🟩
*Web 🟪



Workflow System

Viewing systems exist (e.g., gbrowse, Apollo, Manatee, Artemis), but how to *create* data?

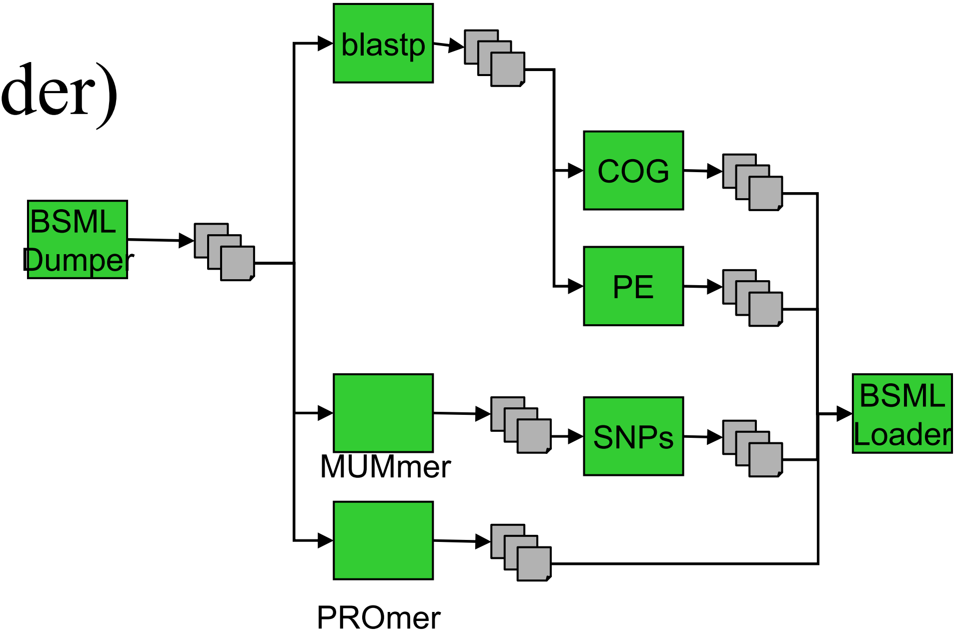
- Implemented in Java, ~5000 lines
- Describes workflow as a directed acyclic graph
- Supports serial and parallel processes
- Executes onto Condor/LSF
- Two main files:
  - Human-readable config files
  - XML templates
- config file + XML template → XML "instance"

Workflow System

- XML instance, contents:
  - Complete description of pipeline
  - Contains status of pipeline
  - Allows monitoring:
    - resumption of failed instances
    - straightforward tracking of multiple instances

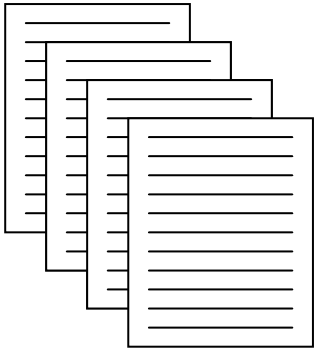
Workflow Computes

- Blast
- Position effect (conserved gene order)
- MUMmer
  - SNPs
- PROmer
- Gene families
  - COGs
  - Paralogs

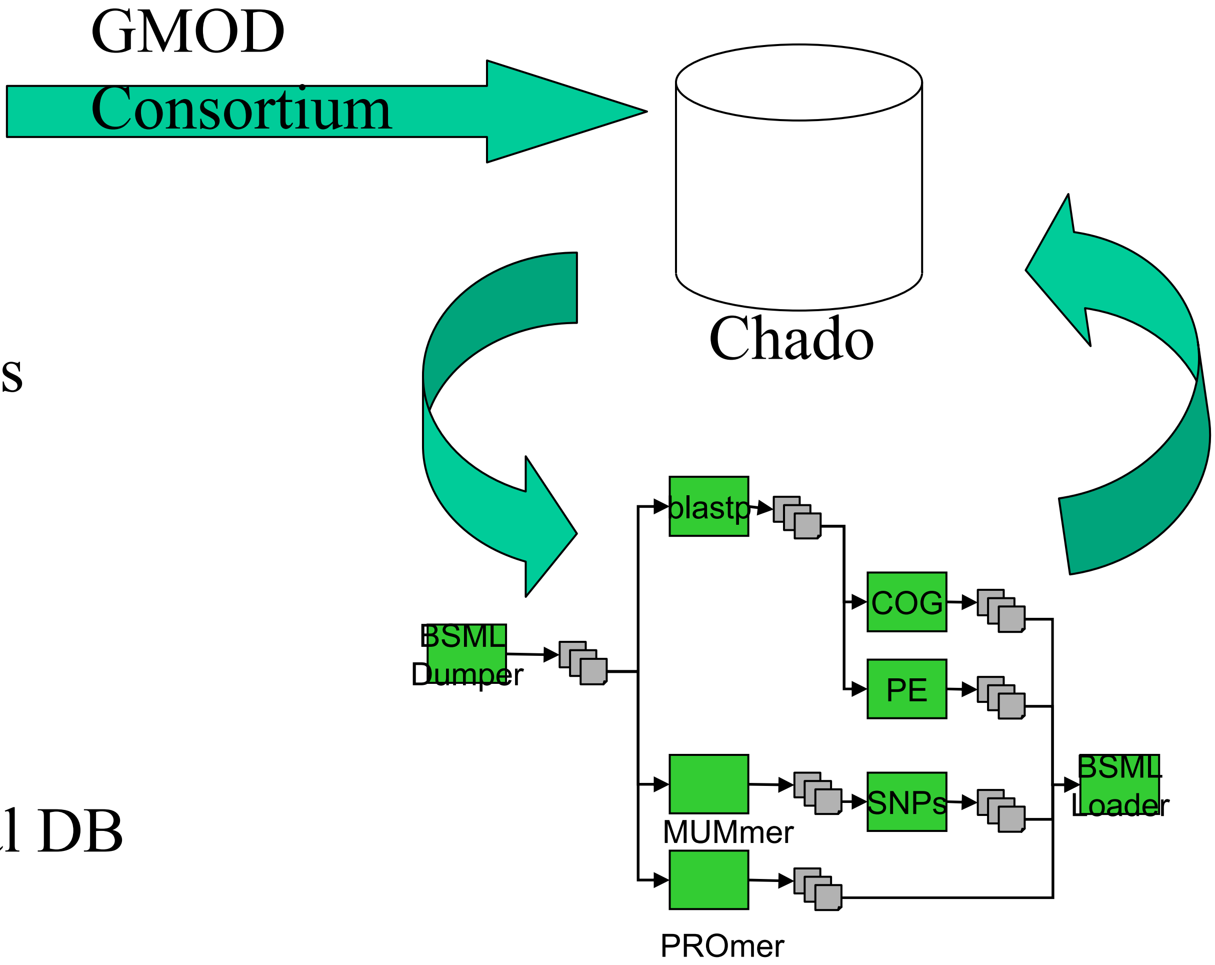


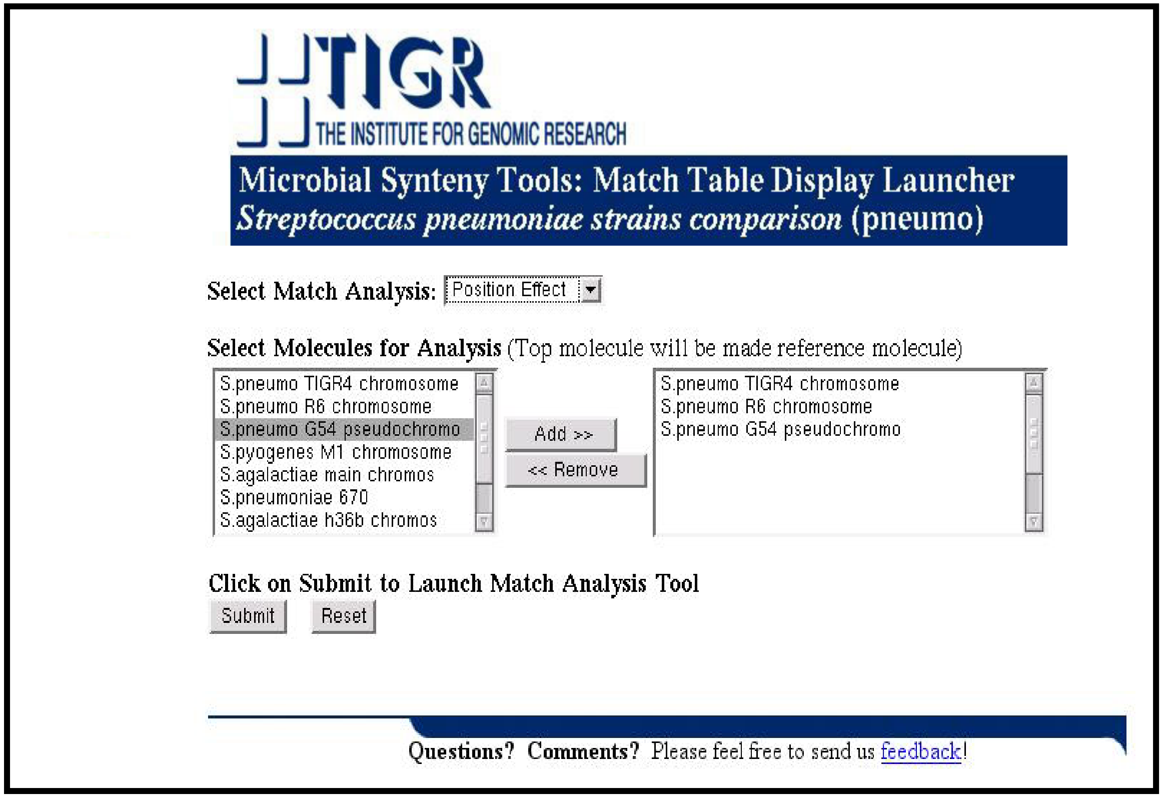
Primary output: BSML-XML

Data Prep For Comparative Analysis



GenBank Files
EMBL Files
Custom Files
XML
    -GFF
    -BSML
TIGR Original DB





TIGR THE INSTITUTE FOR GENOMIC RESEARCH

Microbial Synteny Tools: Match Table Display Launcher
Streptococcus pneumoniae strains comparison **(pneumo)**

Select Match Analysis: Position Effect

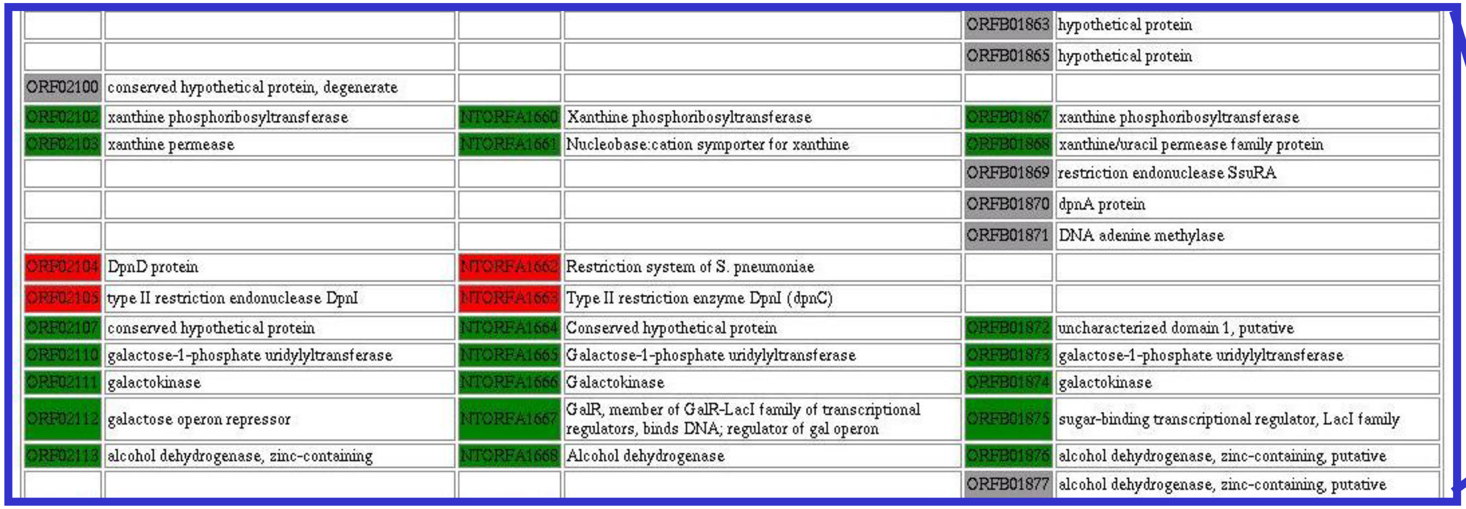
Select Molecules for Analysis (Top molecule will be made reference molecule)

S.pneumo TIGR4 chromosome
S.pneumo R6 chromosome
S.pneumo G54 pseudochromo
S.pyogenes M1 chromosome
S.agalactiae main chromos
S.pneumoniae 670
S.agalactiae h36b chromos

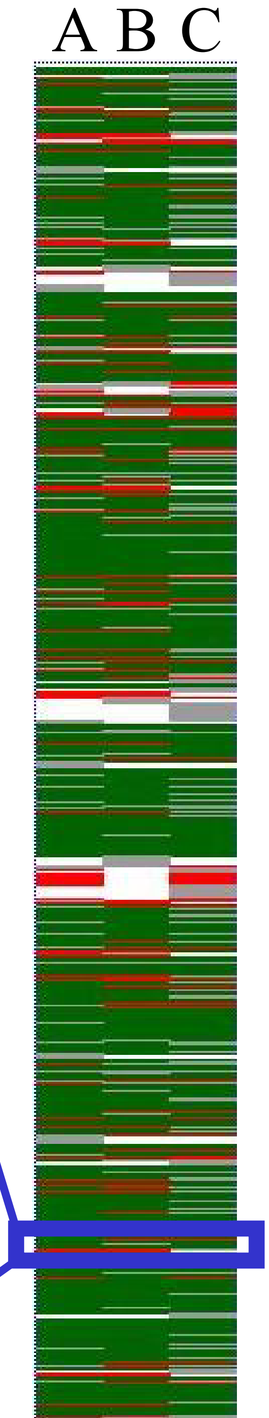
S.pneumo TIGR4 chromosome
S.pneumo R6 chromosome
S.pneumo G54 pseudochromo

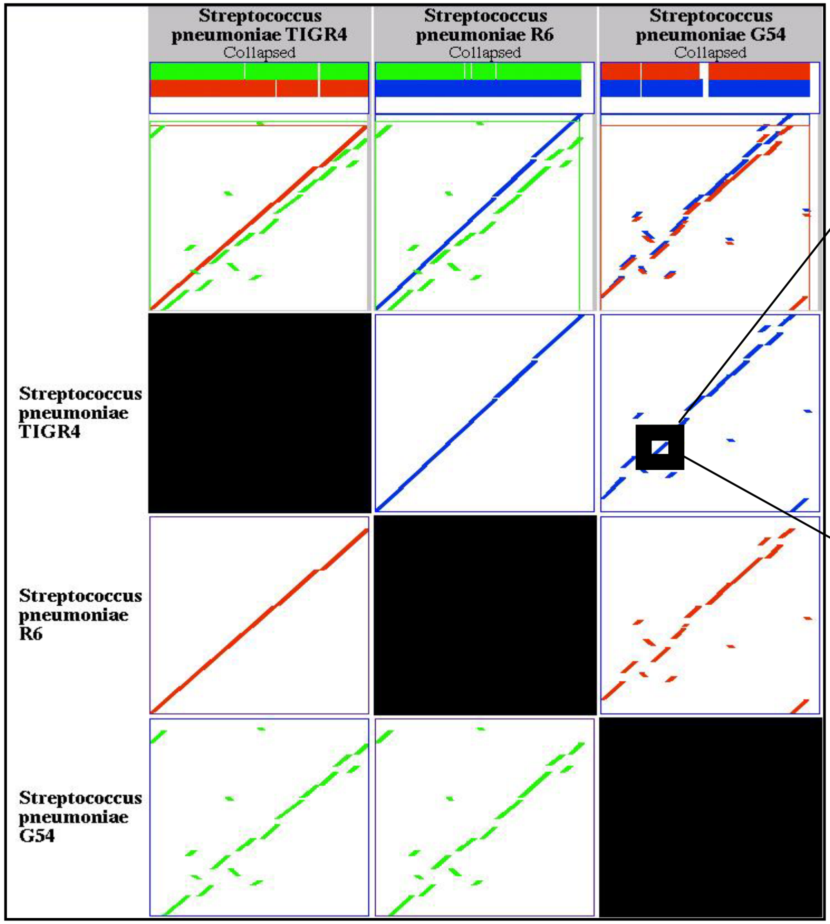
Click on Submit to Launch Match Analysis Tool

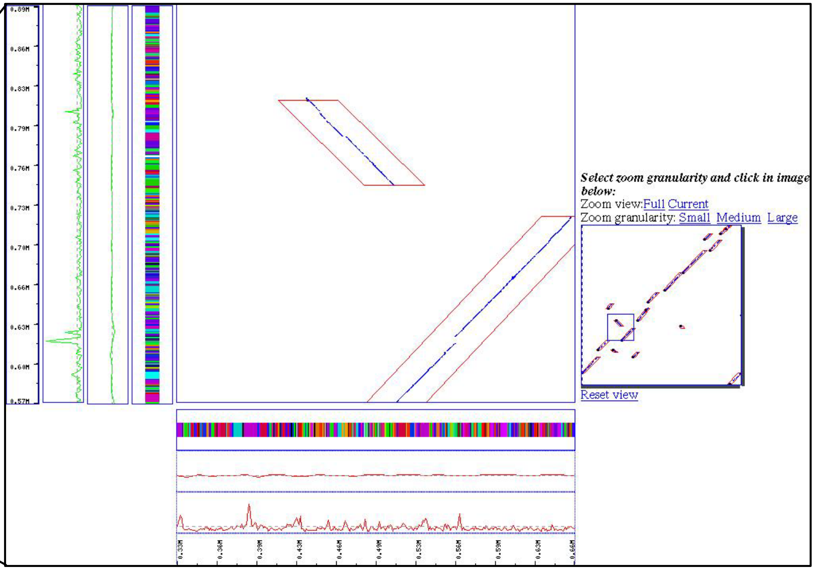
Questions? Comments? Please feel free to send us feedback!



				ORFB01863	hypothetical protein
				ORFB01865	hypothetical protein
ORF02100	conserved hypothetical protein, degenerate				
ORF02102	xanthine phosphoribosyltransferase	NTORFA1660	Xanthine phosphoribosyltransferase	ORFB01867	xanthine phosphoribosyltransferase
ORF02103	xanthine permease	NTORFA1661	Nucleobase:cation symporter for xanthine	ORFB01868	xanthine/uracil permease family protein
				ORFB01869	restriction endonuclease SsuRA
				ORFB01870	dpnA protein
				ORFB01871	DNA adenine methylase
ORF02104	DpnD protein	NTORFA1662	Restriction system of S. pneumoniae		
ORF02105	type II restriction endonuclease DpnI	NTORFA1663	Type II restriction enzyme DpnI (dpnC)		
ORF02107	conserved hypothetical protein	NTORFA1664	Conserved hypothetical protein	ORFB01872	uncharacterized domain 1, putative
ORF02110	galactose-1-phosphate uridylyltransferase	NTORFA1665	Galactose-1-phosphate uridylyltransferase	ORFB01873	galactose-1-phosphate uridylyltransferase
ORF02111	galactokinase	NTORFA1666	Galactokinase	ORFB01874	galactokinase
ORF02112	galactose operon repressor	NTORFA1667	GalR, member of GalR-LacI family of transcriptional regulators, binds DNA; regulator of gal operon	ORFB01875	sugar-binding transcriptional regulator, LacI family
ORF02113	alcohol dehydrogenase, zinc-containing	NTORFA1668	Alcohol dehydrogenase	ORFB01876	alcohol dehydrogenase, zinc-containing, putative
				ORFB01877	alcohol dehydrogenase, zinc-containing, putative





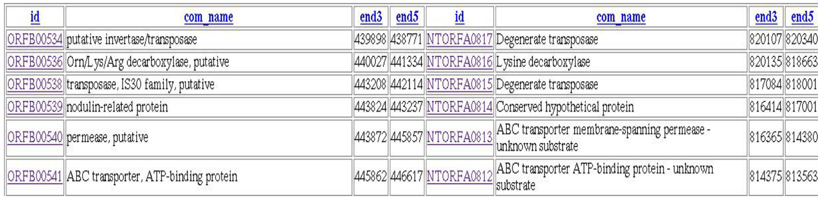


Select zoom granularity and click in image below:

Zoom view:Full Current

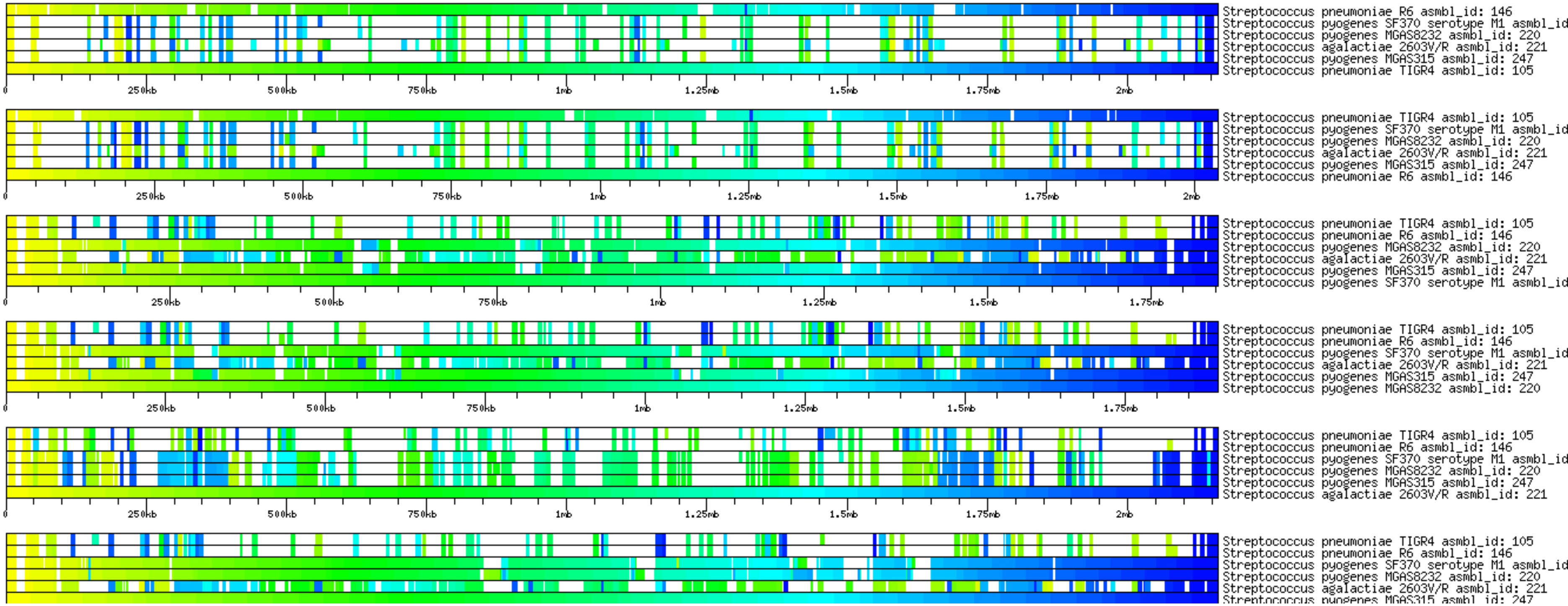
Zoom granularity: Small Medium Large

Reset view

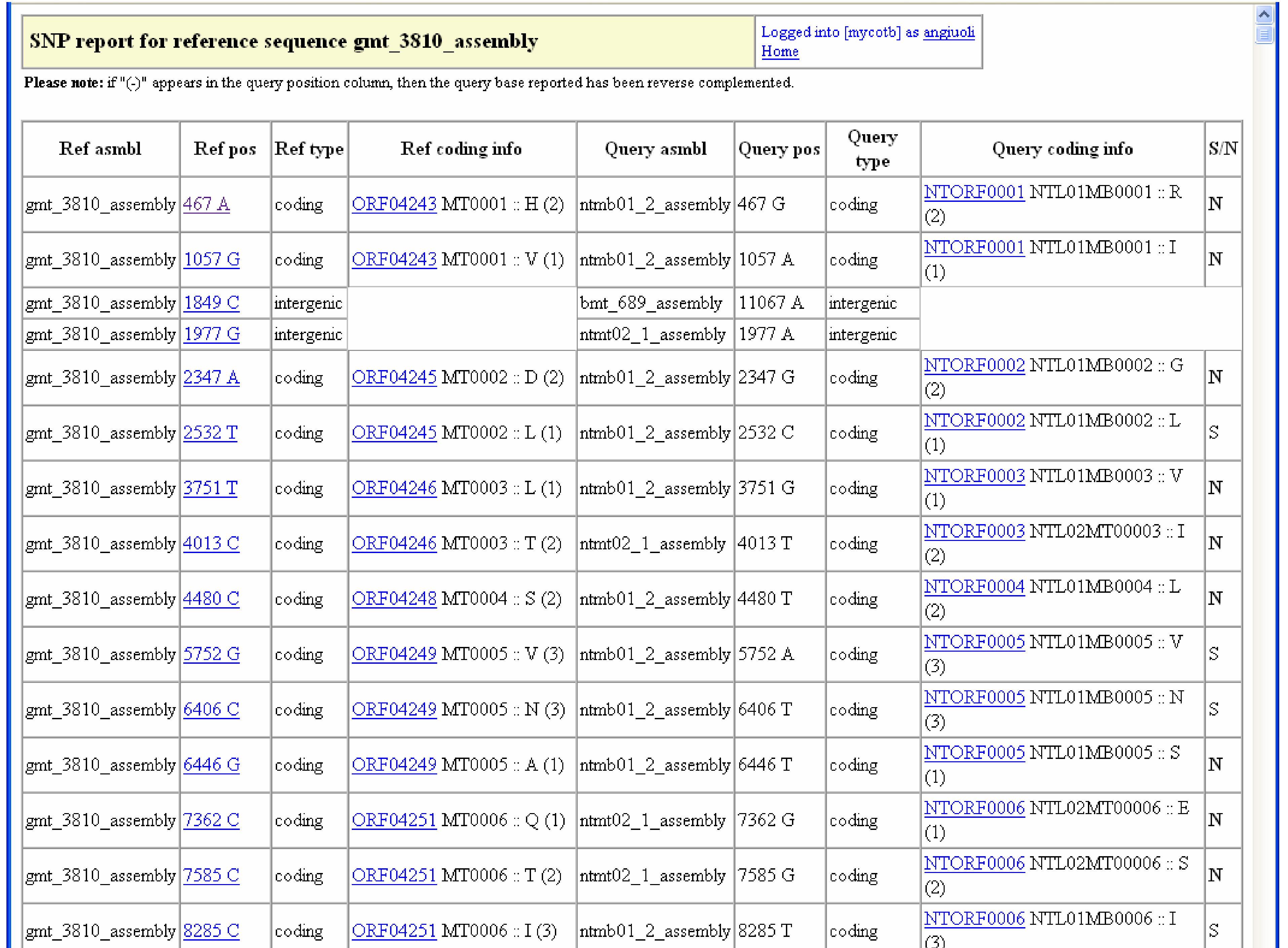


id	com_name	end3	end5	id	com_name	end3	end5
ORFB00534	putative invertase/transposase	439898	438771	NTORFA0817	Degenerate transposase	820107	820340
ORFB00536	Orn/Lys/Arg decarboxylase, putative	440027	441334	NTORFA0816	Lysine decarboxylase	820135	818663
ORFB00538	transposase, IS30 family, putative	443208	442114	NTORFA0815	Degenerate transposase	817084	818001
ORFB00539	nodulin-related protein	443824	443237	NTORFA0814	Conserved hypothetical protein	816414	817001
ORFB00540	permease, putative	443872	445857	NTORFA0813	ABC transporter membrane-spanning permease - unknown substrate	816365	814380
ORFB00541	ABC transporter, ATP-binding protein	445862	446617	NTORFA0812	ABC transporter ATP-binding protein - unknown substrate	814375	813563

XY plots



Conserved synteny across multiple genomes.
Color highlighting of rearrangements at
whole genome level of resolution

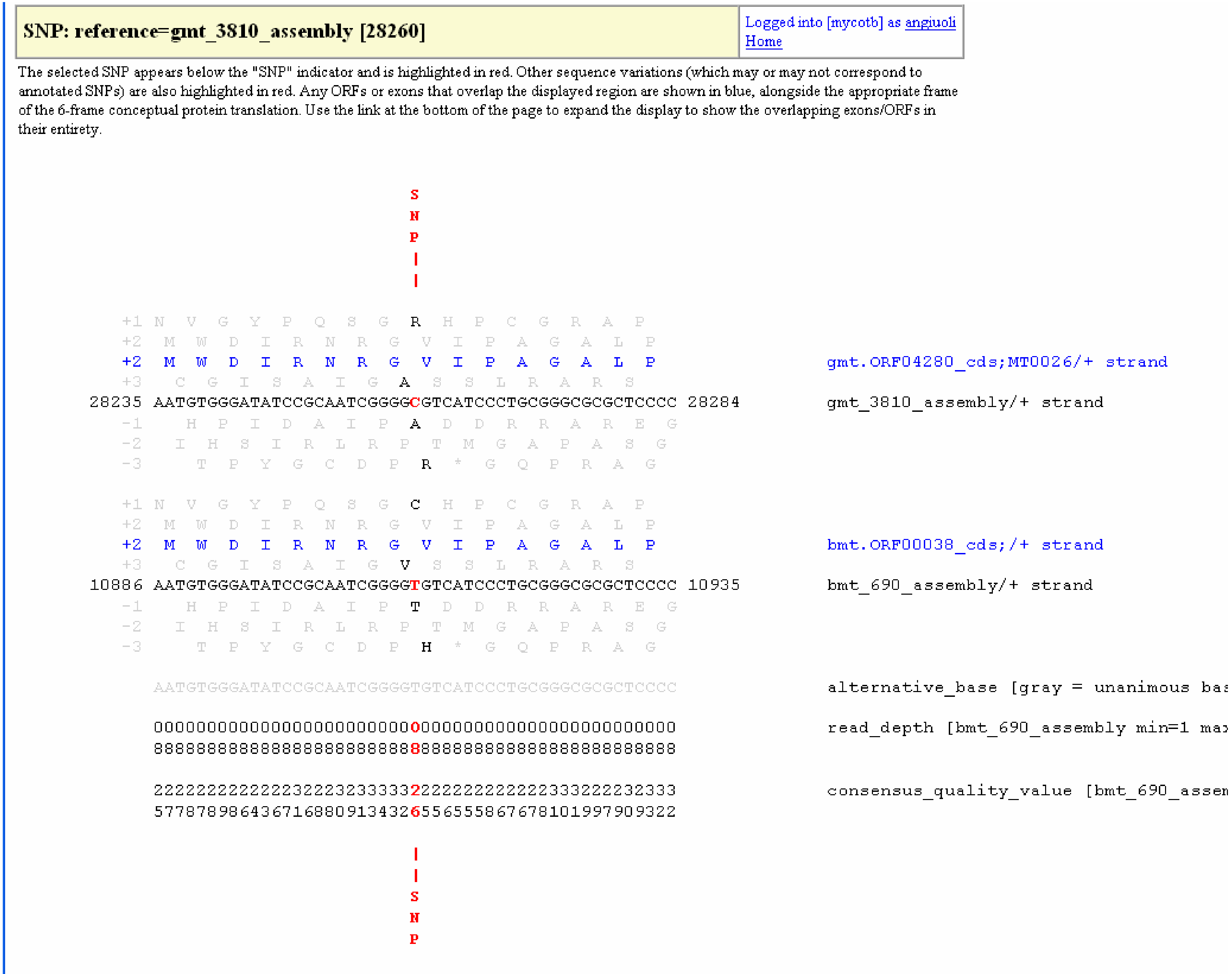


SNP report for reference sequence gmt_3810_assembly

Logged into [mycotb] as angiuoli
Home

Please note: if "(-)" appears in the query position column, then the query base reported has been reverse complemented.

Ref asmbl	Ref pos	Ref type	Ref coding info	Query asmbl	Query pos	Query type	Query coding info	S/N
gmt_3810_assembly	467 A	coding	ORF04243 MT0001 :: H (2)	ntmb01_2_assembly	467 G	coding	NTORF0001 NTL01MB0001 :: R (2)	N
gmt_3810_assembly	1057 G	coding	ORF04243 MT0001 :: V (1)	ntmb01_2_assembly	1057 A	coding	NTORF0001 NTL01MB0001 :: I (1)	N
gmt_3810_assembly	1849 C	intergenic		bmt_689_assembly	11067 A	intergenic		
gmt_3810_assembly	1977 G	intergenic		ntmt02_1_assembly	1977 A	intergenic		
gmt_3810_assembly	2347 A	coding	ORF04245 MT0002 :: D (2)	ntmb01_2_assembly	2347 G	coding	NTORF0002 NTL01MB0002 :: G (2)	N
gmt_3810_assembly	2532 T	coding	ORF04245 MT0002 :: L (1)	ntmb01_2_assembly	2532 C	coding	NTORF0002 NTL01MB0002 :: L (1)	S
gmt_3810_assembly	3751 T	coding	ORF04246 MT0003 :: L (1)	ntmb01_2_assembly	3751 G	coding	NTORF0003 NTL01MB0003 :: V (1)	N
gmt_3810_assembly	4013 C	coding	ORF04246 MT0003 :: T (2)	ntmt02_1_assembly	4013 T	coding	NTORF0003 NTL02MT00003 :: I (2)	N
gmt_3810_assembly	4480 C	coding	ORF04248 MT0004 :: S (2)	ntmb01_2_assembly	4480 T	coding	NTORF0004 NTL01MB0004 :: L (2)	N
gmt_3810_assembly	5752 G	coding	ORF04249 MT0005 :: V (3)	ntmb01_2_assembly	5752 A	coding	NTORF0005 NTL01MB0005 :: V (3)	S
gmt_3810_assembly	6406 C	coding	ORF04249 MT0005 :: N (3)	ntmb01_2_assembly	6406 T	coding	NTORF0005 NTL01MB0005 :: N (3)	S
gmt_3810_assembly	6446 G	coding	ORF04249 MT0005 :: A (1)	ntmb01_2_assembly	6446 T	coding	NTORF0005 NTL01MB0005 :: S (1)	N
gmt_3810_assembly	7362 C	coding	ORF04251 MT0006 :: Q (1)	ntmt02_1_assembly	7362 G	coding	NTORF0006 NTL02MT00006 :: E (1)	N
gmt_3810_assembly	7585 C	coding	ORF04251 MT0006 :: T (2)	ntmt02_1_assembly	7585 G	coding	NTORF0006 NTL02MT00006 :: S (2)	N
gmt_3810_assembly	8285 C	coding	ORF04251 MT0006 :: I (3)	ntmb01_2_assembly	8285 T	coding	NTORF0006 NTL01MB0006 :: I (3)	S



SNP: reference=gmt_3810_assembly [28260]

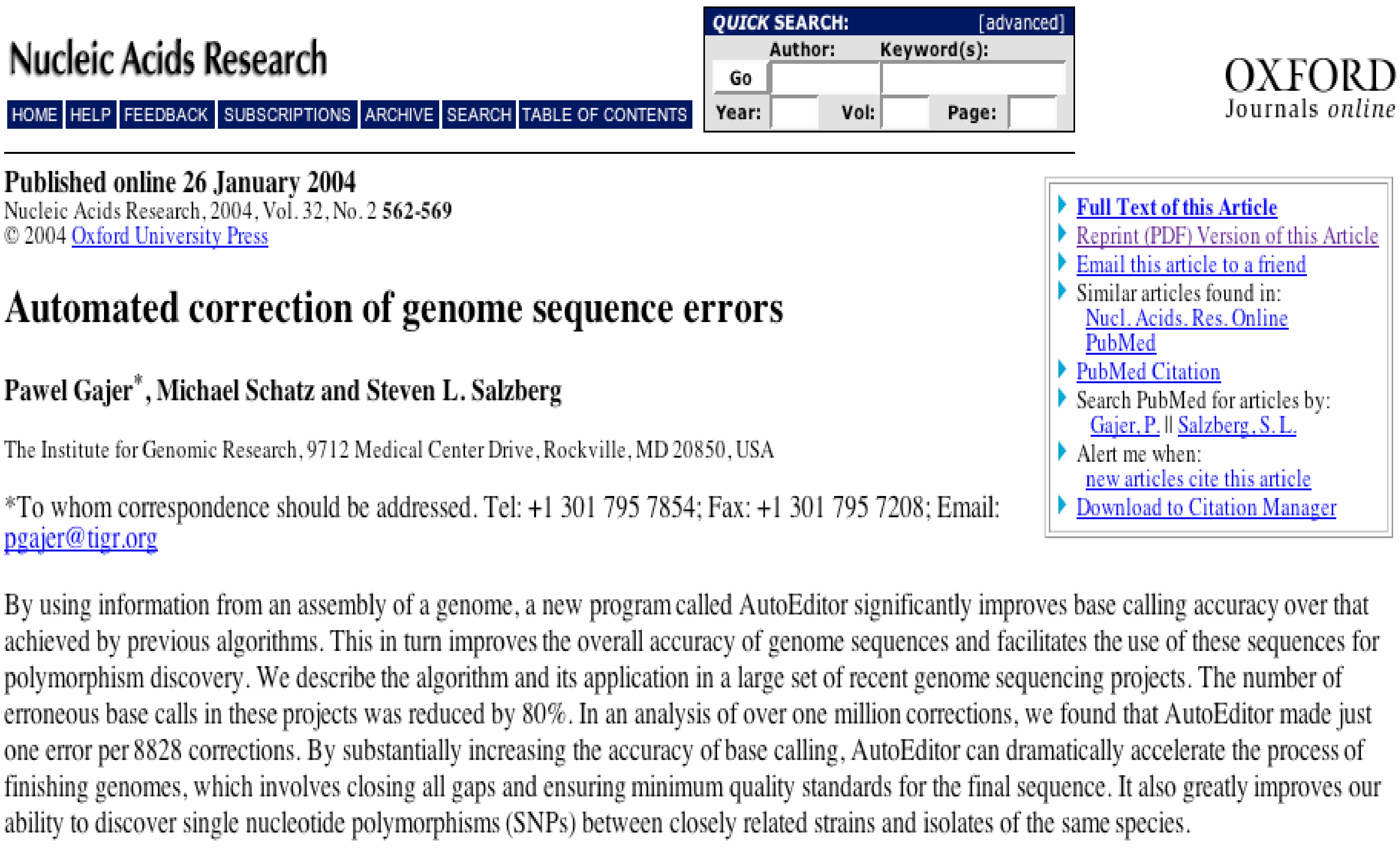
Logged into [mycotb] as angiuoli
Home

The selected SNP appears below the "SNP" indicator and is highlighted in red. Other sequence variations (which may or may not correspond to annotated SNPs) are also highlighted in red. Any ORFs or exons that overlap the displayed region are shown in blue, alongside the appropriate frame of the 6-frame conceptual protein translation. Use the link at the bottom of the page to expand the display to show the overlapping exons/ORFs in their entirety.

```
                                 S
                                 N
                                 P
                                 |
                                 |

      +1 N  V  G  Y  P  Q  S  G  R  H  P  C  G  R  A  P
      +2  M  W  D  I  R  N  R  G  V  I  P  A  G  A  L  P
      +2  M  W  D  I  R  N  R  G  V  I  P  A  G  A  L  P          gmt.ORF04280_cds;MT0026/+ strand
      +3   C  G  I  S  A  I  G  A  S  S  L  R  A  R  S
   28235 AATGTGGGATATCCGCAATCGGGGCGTCATCCCTGCGGGCGCGCTCCCC 28284     gmt_3810_assembly/+ strand
      -1    H  P  I  D  A  I  P  A  D  D  R  R  A  R  E  G
      -2   I  H  S  I  R  L  R  P  T  M  G  A  P  A  S  G
      -3     T  P  Y  G  C  D  P  R  *  G  Q  P  R  A  G

      +1 N  V  G  Y  P  Q  S  G  C  H  P  C  G  R  A  P
      +2  M  W  D  I  R  N  R  G  V  I  P  A  G  A  L  P
      +2  M  W  D  I  R  N  R  G  V  I  P  A  G  A  L  P          bmt.ORF00038_cds;/+ strand
      +3   C  G  I  S  A  I  G  V  S  S  L  R  A  R  S
   10886 AATGTGGGATATCCGCAATCGGGGTGTCATCCCTGCGGGCGCGCTCCCC 10935     bmt_690_assembly/+ strand
      -1    H  P  I  D  A  I  P  T  D  D  R  R  A  R  E  G
      -2   I  H  S  I  R  L  R  P  T  M  G  A  P  A  S  G
      -3     T  P  Y  G  C  D  P  H  *  G  Q  P  R  A  G

         AATGTGGGATATCCGCAATCGGGGTGTCATCCCTGCGGGCGCGCTCCCC           alternative_base [gray = unanimous ba:

         0000000000000000000000000000000000000000000000000           read_depth [bmt_690_assembly min=1 ma:
         8888888888888888888888888888888888888888888888888

         2222222222222322232333332222222222222333222232333           consensus_quality_value [bmt_690_assen
         5778789864367168809134326556555867678101997909322

                                 |
                                 |
                                 S
                                 N
                                 P
```

Sybil Completion

- Data model – 100%
- Workflow process – 100%
- Loading systems – 100%
- Interfaces – 80%, on-going
- SNP analysis – early phase
- Higher-level analysis systems – early phase



Nucleic Acids Research
HOME
HELP
FEEDBACK
SUBSCRIPTIONS
ARCHIVE
SEARCH
TABLE OF CONTENTS
QUICK SEARCH:
[advanced]
Author:
Keyword(s):
Go
Year:
Vol:
Page:
OXFORD
Journals online

Published online 26 January 2004
Nucleic Acids Research, 2004, Vol. 32, No. 2 **562-569**
© 2004 Oxford University Press

Full Text of this Article
Reprint (PDF) Version of this Article
Email this article to a friend
Similar articles found in:
Nucl. Acids. Res. Online
PubMed
PubMed Citation
Search PubMed for articles by:
Gajer, P. || Salzberg, S. L.
Alert me when:
new articles cite this article
Download to Citation Manager

Automated correction of genome sequence errors

Pawel Gajer*, Michael Schatz and Steven L. Salzberg

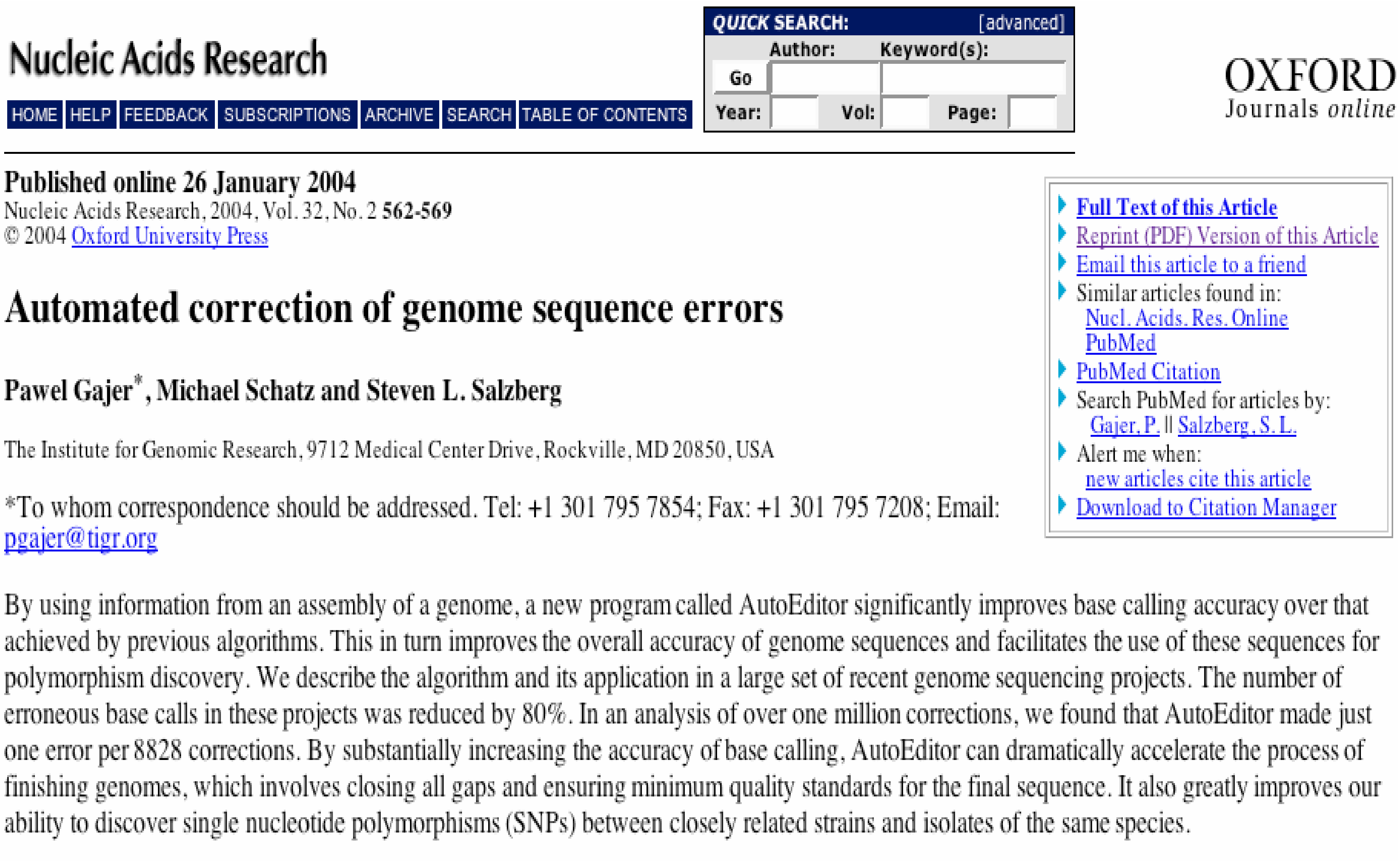
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

*To whom correspondence should be addressed. Tel: +1 301 795 7854; Fax: +1 301 795 7208; Email: pgajer@tigr.org

By using information from an assembly of a genome, a new program called AutoEditor significantly improves base calling accuracy over that achieved by previous algorithms. This in turn improves the overall accuracy of genome sequences and facilitates the use of these sequences for polymorphism discovery. We describe the algorithm and its application in a large set of recent genome sequencing projects. The number of erroneous base calls in these projects was reduced by 80%. In an analysis of over one million corrections, we found that AutoEditor made just one error per 8828 corrections. By substantially increasing the accuracy of base calling, AutoEditor can dramatically accelerate the process of finishing genomes, which involves closing all gaps and ensuring minimum quality standards for the final sequence. It also greatly improves our ability to discover single nucleotide polymorphisms (SNPs) between closely related strains and isolates of the same species.

SNP Detection Pipeline Output

```
>KRUGERB-16 (94989bp) 200731 c-->t 50878 COV: 16 CB_QVal: 504 QUAL: 31.5
NON-SYNONYMOUS (2): GGC - G ---> GAC - D   (from start: 4 bp/848 bp)
REF : GAAGAAGCGCTAATAAAATGCCCATTACAAGACTCCCTTCG 200751 B. anthracis Ames Porton
      |||||||||||||||||||| ||||||||||||||||||||
ASM : GAAGAAGCGCTAATAAAATGTCCATTACAAGACTCCCTTCG 50898
ORF01519        transporter  putative
 50873   A       564      AAAAAAAAAAAAAAAA   35:36:35:35:36:36:36:36:35:27:36:36:36:36:37:36
 50874   A       569      AAAAAAAAAAAAAAAA   34:36:35:35:36:36:36:36:35:35:36:36:36:36:35:36
 50875   A       550      AAAAAAAAAAAAAAAA   18:36:35:35:36:36:36:36:34:36:36:35:36:34:35:36
 50876   T       545      TTTTTTTTTTTTTTTT   13:36:36:35:36:36:36:34:35:38:35:34:36:34:37:34
 50877   G       517      GGGGGGGGGGGGGGGG   17:35:35:35:34:37:36:36:32:26:34:34:33:36:22:35
*50878   T       504      TTTTTTTTTTTTTTTT   11:36:35:33:34:36:31:34:35:19:31:35:31:36:33:34
 50879   C       559      CCCCCCCCCCCCCCCC   33:34:35:35:36:36:34:35:32:35:36:35:36:36:35:36
 50880   C       515      CCCCCCCCCCCCCCC    34:35:35:36:34:36:36:29:34:36:35:35:36:35:29
 50881   A       508      AAAAAAAAAAAAAAA    36:34:34:36:40:34:36:32:26:36:34:36:36:32:26
 50882   T       513      TTTTTTTTTTTTTTT    36:35:35:36:35:36:36:36:25:36:37:35:36:29:30
 50883   T       498      TTTTTTTTTTTTTTT    36:35:34:36:33:36:36:35:15:36:35:32:36:33:30
REFERENCE: B. anthracis Ames Porton
 200726        A       334      AAAAAAAAAA   29:29:37:31:51:22:45:40:35:15
 200727        A       362      AAAAAAAAAA   25:41:37:34:51:27:45:40:34:28
 200728        A       347      AAAAAAAAAA   29:41:37:34:45:27:45:40:34:15
 200729        T       341      TTTTTTTTTT   25:41:45:33:45:21:45:25:34:27
 200730        G       336      GGGGGGGGGG   33:41:38:33:45:16:45:24:34:27
*200731        C       334      CCCCCCCCCC   33:37:40:37:45:16:18:45:35:28
 200732        C       318      CCCCCCCCCC   33:37:40:38:45:18:4:45:35:23
 200733        C       284      CCCCCCCCC    33:37:37:37:41:15:4:45:35
 200734        A       307      AAAAAAAAA    27:45:37:37:51:21:4:51:34
 200735        T       291      TTTTTTTTT    24:45:37:37:41:24:4:45:34
 200736        T       293      TTTTTTATT    24:45:45:37:41:24:4:45:35
COV: 10 CBQUAL: 334 QUAL: 33.4
```


QUICK SEARCH:
[advanced]
Author:
Keyword(s):
Go
Year:
Vol:
Page:
HOME
HELP
FEEDBACK
SUBSCRIPTIONS
ARCHIVE
SEARCH
TABLE OF CONTENTS
OXFORD
Journals online
Full Text of this Article
Reprint (PDF) Version of this Article
Email this article to a friend
Similar articles found in:
Nucl. Acids. Res. Online
PubMed
PubMed Citation
Search PubMed for articles by:
Gajer, P. || Salzberg, S. L.
Alert me when:
new articles cite this article
Download to Citation Manager

Nucleic Acids Research

Published online 26 January 2004
Nucleic Acids Research, 2004, Vol. 32, No. 2 **562-569**

Automated correction of genome sequence errors

Pawel Gajer*, Michael Schatz and Steven L. Salzberg

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

*To whom correspondence should be addressed. Tel: +1 301 795 7854; Fax: +1 301 795 7208; Email: pgajer@tigr.org

By using information from an assembly of a genome, a new program called AutoEditor significantly improves base calling accuracy over that achieved by previous algorithms. This in turn improves the overall accuracy of genome sequences and facilitates the use of these sequences for polymorphism discovery. We describe the algorithm and its application in a large set of recent genome sequencing projects. The number of erroneous base calls in these projects was reduced by 80%. In an analysis of over one million corrections, we found that AutoEditor made just one error per 8828 corrections. By substantially increasing the accuracy of base calling, AutoEditor can dramatically accelerate the process of finishing genomes, which involves closing all gaps and ensuring minimum quality standards for the final sequence. It also greatly improves our ability to discover single nucleotide polymorphisms (SNPs) between closely related strains and isolates of the same species.